Characterization of Protein Complexes and Protein Interaction Networks by Mass Spectrometry

Dissertation

zur Erlangung des akademischen Grades Doktor rerum naturalium (Dr. rer. nat)

Fakultät Mathematik und Naturwissenschaften
Technische Universität Dresden

Dipl. Biol. Anna Shevchenko

aus Sankt Petersburg, Russland

Gutachter:

Prof. Dr. Hoflack, TU Dresden

Prof. Dr. Rödel, TU Dresden

Prof. Dr. Jensen, Universität of Southern Denmark

Tag der Einreichung: 27.07.2004

Datum der Verteidigung: 22.11.2004

Part of this work was published in:

Chapter 3.1

Shevchenko, A., Shevchenko, A. (2001) Evaluation of the efficiency of in-gel digestion of proteins by peptide isotopic labeling and MALDI mass spectrometry. *Anal Biochem* **296**, 279-283.

Shevchenko, A., Loboda, A., Ens, W., Schraven, B., Standing, K.G., Shevchenko, A. (2001) Archived polyarylamide gels as a resource for proteome characterization by mass spectrometry. *Electrophoresis* **22**, 1194-1203.

Shevchenko, A., Sunyaev, S., Liska, A., Bork, P and Shevchenko, A. (2003). Nanoelectrospray tandem mass spectrometry and sequence similarity searching for identification of proteins from organisms with unknown genomes. *Methods Mol Biol* 211, 221-234.

Chapter 3.2

Rigaut, G., **Shevchenko, A**., Rutz, B., Wilm, M., Mann, M., Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**, 1030-1032.

Shevchenko, A., Schaft, D., Roguev, A., Pijnappel, W.W.M.P., Stewart, A.F., Shevchenko, A. (2002) Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry: analytical perspective. *Mol Cell Proteomics* **1**, 204-212.

Chapter 3.3

Shevchenko, A., Zachariae, W., Shevchenko, A. (1999) A strategy for the characterization of protein interaction networks by mass spectrometry. *Biochem. Soc. Trans*. **27**, 549-554.7.

Roguev*, A., **Shevchenko*, A**., Schaft, D., Thomas, H., Stewart, A.F., Shevchenko, A. (2004) A comparative analysis of an orthologous proteomic environment in the yeasts *S.cerevisiae* and *S. pombe*. *Mol Cell Proteomics* **3**, 125-132.
* equal contribution

The characterization of protein complexes described in Chapters 3.2 and 3.3 was performed in collaboration projects with groups of Dr. B. Seraphin (EMBL, Heidelberg) and Prof.Dr. A.F.Stewart (EMBL, Heidelberg; BIOTEC TU Dresden) and published in the following papers:

Bouveret, E., Rigaut, G., **Shevchenko, A**., Wilm, M., and Seraphin, B. (2000). A Sm-like protein complex that participates in mRNA degradation. *EMBO J* **19**, 1661-1671.

Caspary, F., **Shevchenko, A**., Wilm, M., and Seraphin, B. (1999). Partial purification of the yeast U2 snRNP reveals a novel yeast pre- mRNA splicing factor required for pre-spliceosome assembly. *EMBO J* **18**, 3463-3474.

Roguev, A., Schaft, D., **Shevchenko, A.,** Pijnappel, W. W. M., Wilm, M., Aasland, R., and Stewart, A. F. (2001). The *S. cerevisiae* Set1 complex includes an Ash2 homolog and methylates histone 3 lysine 4. *EMBO J* **20**, 7137-7148.

Pijnappel, W. W., Schaft, D., Roguev, A., **Shevchenko,** A., Tekotte, H., Wilm, M., Rigaut, G., Seraphin, B., Aasland, R., and Stewart, A. F. (2001). The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes Dev* **15**, 2991-3004.

Roguev, A., Schaft, D., **Shevchenko, A.**, Aasland, R., and Stewart, A. F. (2003). High conservation of the Set1/Rad6 axis of histone 3 lysine 4 methylation in budding and fission yeasts. *J Biol Chem* **278**, 8487-8493.

Schaft, D., Roguev, A., Kotovic, K. M., **Shevchenko, A**., Sarov, M., Shevchenko, A., Neugebauer, K. M., and Stewart, A. F. (2003). The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation. *Nucleic Acid Res* **31**, 2475-2482

List of abbreviations

AA/BIS	Acrylamide/bisacrylamide
APS	Ammonium persulfate
BSA	Bovine serum albumin
CAI	Codon adaption index
CBI	Codon bias index
CBP	Calmodulin-binding peptide
CHCA	α-cyano-4-hydroxy-*trans*-cinnamic acid
C-terminus	Carboxyl terminus
DHB	2,5-Dihydroxybenzoic acid
DNA	Desoxyribonucleic acid
DTT	Dithiothreitol
ESI	Electrospray ionization
EGTA	Ethyleneglycol-O, O'-bis(2-aminoethyl)-N, N, N', N'-tetraacetic acid
FT ICR	Fourier transform ion cyclotron resonance
GFP	[Glu]-fibrinogen peptide
HPLC	High performance liquid chromatigraphy
IP	Immunoprecipitation
IgG	Immunoglobulin G
kDa	kiloDalton
MALDI	Matrix-assisted laser desorption/ionization
MS	Mass spectrometry
MS BLAST	Mass spectrometry driven BLAST
MW	Molecular weight
MudPIT	MultiDimensional Protein Identification Technique
NAD	Nicotinamide adenine dinucleotide
nanoES	Nanoelectrospray
N-terminus	Amino terminus
ns	Nanoseconds
oMALDI-qQTOF	Orthogonal MALDI quadrupole time-of-flight
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PMF	Peptide mass fingerprint
PombePD	*S. pombe* protein database
ProtA	Protein A
TEMED	N,N,N',N'-tetramethylenediamine
QqTOF	Quadrupole time-of-flight
RF quadrupole	Radio-frequency quadrupole
SEAM	SEquential rounds of immunoAffinity purifications/Mass spectrometry
SDS	Sodium dodecylsulfate
TAP	Tandem affinity purification
TEV	Tobacco etch virus
TFA	Trifluoroacetic acid
TOF	Time-of-flight
Tris	Tris(hydroxymethyl)aminomethane
v/v	Volume per volume
w/v	Weight per volume

Y2H	Yeast two hybrid
YPD	Yeast protein database

Amino acids (one-letter code)

A	alanine	M	methionine
C	cysteine	N	asparagine
D	aspartic acid	P	proline
E	glutamic acid	Q	glutamine
F	phenylalanine	R	arginine
G	glycine	S	serine
H	histidine	T	threonine
I	isoleucine	V	valine
K	lysine	W	tryptophan
L	leucine	Y	tyrosine

Contents

1. Introduction

1.1 Mass spectrometry as an analytical tool for proteomics

Various technical disciplines, including cell imaging by light and electron microscopy, array and chip experiments as well as genetic readout experiments, such as the yeast two-hybrid screens, contribute to the field of proteomics. Mass spectrometry has become an indispensable tool in molecular and cellular biology because of its ability to identify and to quantify thousands of proteins in complex biological samples. Mass spectrometry played a key role in charting of protein–protein interactions at the proteome-wide scale [1, 2], characterization of the protein composition of cellular organelles [3-9], the concurrent description of the malaria parasite genome and proteome [10, 11] as well as in the quantitative characterization of protein profiles in membrane microdomains [12, 13].

Mass spectrometry-based proteomics is a synthetic discipline, which utilizes developments in genomic sequencing, molecular and technical physics and bioinformatics. The first section of this chapter will discuss recent innovations in mass spectrometry instrumentation and technology.

1.1.1 Mass spectrometric instrumentation for proteomic applications

1.1.1.1 Overview of ionization methods and mass analyzers

Mass spectrometric measurements are carried out on ionized analytes in the gas phase. A typical mass spectrometer consists of an ion source, a mass analyzer that discriminates ions of the analyte by their mass-to-charge ratio (m/z), and an ion detector. Matrix-assisted laser desorption/ionization (MALDI, Figure 1A) [14] and electrospray ionization (ESI, Figure 2B) [15] are most common ionization methods employed in the analysis of proteins and peptides. ESI ionizes analytes out of acidified aqueous methanol or acetonitrile, whereas in MALDI laser pulses ablate and ionize analytes that were co- crystallized with UV-absorbing organic acids used as matrices [16-18].

Ions produced in the source are then transported into a mass analyzer. Four major types of mass analyzers are typically employed in proteomics [19, 20]- ion trap, time-of-flight (TOF), quadrupole and Fourier Transform Ion Cyclotron Resonance (FT ICR). They are very different in design and performance, each having own strength and weakness. A mass spectrometer can either comprise a single analyzer (for example, a MALDI TOF instrument), or two analyzers of the same type (a triple quadrupole instrument), or different analyzers can be combined within a hybrid instrument to take advantage of the strengths of each type (a quadrupole time-of-flight or an ion trap FT ICR instruments).

Selection of the most appropriate mass spectrometric technology determines the success of proteomic projects. In this work four different types of mass spectrometers were employed, including a combination of MALDI and ESI ion sources with quadrupole and time-of-flight mass analyzers.

1.1.1.2 MALDI TOF mass spectrometer

MALDI ion source and TOF analyzers are “natural born” partners because of the pulsed nature of laser ionization (Figure 1B). First, protein or peptide analytes are allowed to co-crystallize with organic matrixes on a surface of metal or polymer-coated targets [21]. Co-crystallization is a crucial step in MALDI analysis and is typically achieved by two major methods. In dried-droplet probe preparation [14] aliquots of the sample and the matrix solution are mixed and deposited onto a surface of a stainless steel target. Co-crystallization of the matrix and the analyte occurs during the evaporation of a volatile organic solvent and low molecular weight contaminants, such as salts, are largely excluded from the analyte/matrix crystals. Although these crystals can be rinsed gently with cold diluted TFA, the efficiency of de-salting is usually low. The method is rather simple and lends itself to complete automation, although at the femtomole level the sensitivity and the reproducibility of detected signals are poor. In thin layer probe preparations [22] matrix is first deposited on a stainless steel support and then an acidified aliquot of the analyte is placed on top of a pre-formed matrix layer and allowed to co-crystallize with the matrix. An exposed layer of peptide/matrix crystals can be washed extensively without affecting the quality of TOF spectra. Doping CHCA matrix with nitrocellulose [23] improves the sensitivity by reducing peptide losses that occur

during the washing step. The method provides good sensitivity and mass resolution of MALDI TOF spectra, however it is rather laborious and difficult to automate. Furthermore, so called “sweet spots” – patches of matrix/analyte crystals, from which high quality spectra could be acquired, are randomly distributed at the surface of the target and are difficult to recognize using a video camera built into the ion source. Thin layer crystals are quickly depleted by laser pulses and accurate adjustment of the laser fluence is required during the acquisition of spectra [24].

Matrix/analyte crystals formed at the target are then irradiated by a laser pulse, which generates a short burst of ions. The ions are accelerated to a fixed amount of kinetic energy and travel down a flight tube in a field-free condition. Ions with low *m/z* have higher velocity and are hit the detector before ions with larger *m/z*. Spectra acquired from hundreds of laser shots are usually averaged to produce the final MALDI spectrum. Modern reflector MALDI mass spectrometers typically enable a few parts per million (ppm) mass accuracy, and less than a femtomole of peptide material needs to be deposited onto a target to produce a good quality spectrum.

Because of its simplicity, excellent mass accuracy, high resolution and sensitivity, MALDI TOF MS is a major technology for the high throughput identification of proteins by a method of peptide mass fingerprinting, also referred to as peptide mass mapping (see section 1.1.2.2 for details).

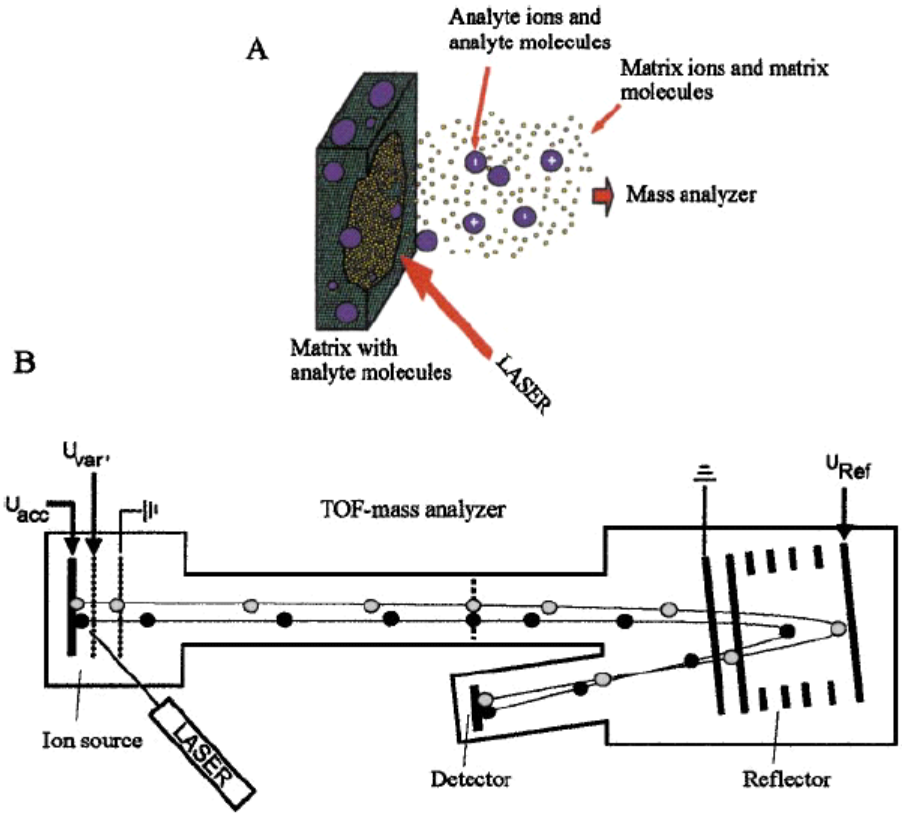


Figure 1. Schematic of MALDI process and TOF instrument. (*A*) A sample co-crystallized with the matrix is irradiated by a laser beam, leading to ablation and ionization of the analyte. (*B*) , Strong acceleration field is switched on 100–500 ns after the laser pulse (delayed extraction), which imparts a fixed kinetic energy to the ions produced by the MALDI process. These ions travel down a flight tube and are turned around by the ion mirror, or reflector, to correct for initial energy differences. The mass-to-charge ratio is related to the time it takes an ion to reach the detector; the ions with smaller *m/z* arriving first. The ions are detected by a channeltron electron multiplier (from [82]).

1.1.1.3 Triple quadrupole tandem mass spectrometer with a NanoES ion source

Quadrupole analyzer is a mass filter, which consists of four parallel rods to which oscillating electric field is applied [25]. At given voltage and frequency, only ions with certain *m/z* can pass through the analyzer, whilst ions with other *m/z* move along unstable trajectories, hit the quadrupole rods, discharge and do not reach the detector (Figure 2C). Mass spectrometer scans *m/z* of ions by changing the amplitude or the frequency of electric field.

A triple quadrupole instrument (Figure 2A), recognized work-horse in many proteomics applications, consists of three sections: two mass-separating quadrupole sections are spaced by a central section with non-scanning RF quadrupole. The non-scanning quadrupole is placed into a collision cell – a differentially pumped chamber, which can be filled with a neutral gas (such as nitrogen or argon). In single MS operational mode the first analytical quadrupole Q1 is scanning, whereas the second analytical quadrupole Q3 and the RF-quadrupole Q2 are only transporting and focusing the ion beam with minimal segregation of ions and the collision cell is not filled with gas. This method is only used to determine the intact masses of analyzed molecules since no fragmentation occurs in the mass spectrometer. However, for many proteomics applications it is desirable to obtain spectra of fragment ions from intact peptide (MS/MS spectra or tandem mass spectra), which assist in determining the peptide sequence *de novo*, or used in database searches, or help to pinpoint post-translational modifications. To this end precursor ions are first selected by Q1 (which operates at the relatively low mass resolution, so that the entire isotopic cluster can be transmitted) and then directed into the collision cell Q2 where they collide with gas molecules and produce fragment ions (details on a quadrupole set-up for peptide sequencing are described in [26]). Q2 prevents fragments from scattering and channels them into the scanning quadrupole Q3, which determines their exact masses.

Quadrupole mass spectrometers are capable of unit mass resolution and mass accuracy of 0.1 – 1 Da and excel at quantitative measurements. Since two quadrupole sections (Q1 and Q3) are operated independently, triple quadrupole machines support a variety of scan modes, which complement basic MS and MS/MS operations. For example, Q3 can be fixed to only transmit ions with the certain characteristic *m/z*, while Q1 is scanning (precursor ion scanning). Ions passed through Q1 would only be detected, if upon their collisional fragmentation they produce the characteristic ion,

which can pass Q3. Precursor ion scanning is vitally important for nano ES MS/MS sequencing or identification of phosphorylated peptides, which relies upon the detection of low intensity target peptides obscured by abundant chemical noise [27, 28].

Miniaturized nano-electrospray sample infusion [29, 30] (Figure 2B) enables extremely low sample consumption because of less than 30 nL/min flow rate. Since new nanoelectrospray needle is used for each analysis, direct injection of unseparated peptide mixtures is possible with no dager of cross-contamination.

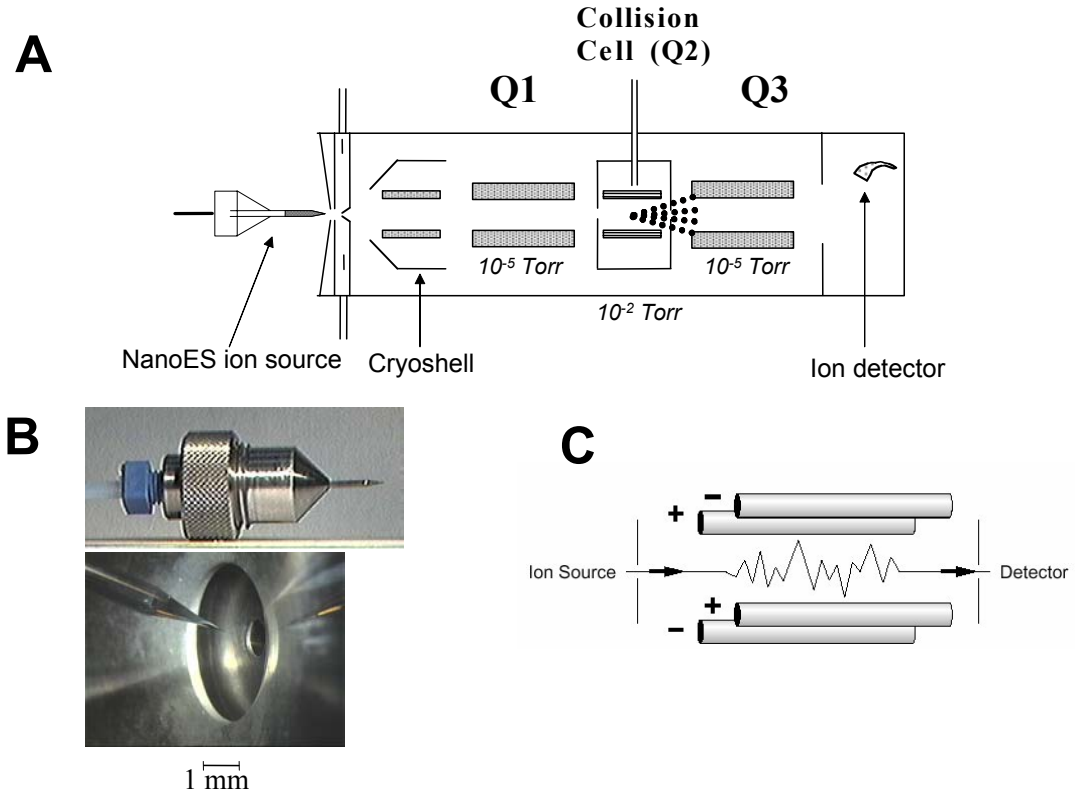


Figure 2. Triple quadrupole tandem mass spectrometer with nanoES ion source. (A). Schematic diagram of the instrument. (B). Miniaturized nanoES source designed at EMBL. Nanoelectrospray capillary shown in front of the orifice of a triple quadrupole mass spectrometer API III contains about 0.5 μL of the analyte solution, which typically enables its spraying for more than 30 min. (C). Schematic diagram of quadrupole mass analyzer (adapted from [30] and [19]).

1.1.1.4 Hybrid quadrupole time-of-flight instruments with nanoES and MALDI ion sources

A hybrid quadrupole TOF tandem mass spectrometer can be described in the simplest way as a triple quadrupole machine with the last quadrupole section replaced by a TOF analyzer [31]. In this work, as well as in the original articles, quadrupole TOF machines are termed as QqTOF, to underscore that that a non-scanning quadrupole q is positioned between the quadrupole and TOF sections inside the collision cell.

In a typical QqTOF design, an additional quadrupole Q0 is placed in front of the analytical quadrupole Q1 to provide collisional damping and focusing of ions, so the instrument (Figure 3A) consists of three quadrupoles – RF-quadrupoles Q0 and Q2 and the analytical quadrupole Q1, followed by a reflecting TOF mass analyzer with orthogonal injection of ions. For single MS (or TOF MS) measurements, the mass filter Q1 is operated as a transmission element, while the TOF analyzer is used to record spectra. For MS/MS, Q1 is operated in the mass filter mode to transmit only selected precursor ions. In precursor ion scanning mode Q1 is scanning, while TOF analyzer acquires the full range spectrum of fragment ions [32].

One of the main advantages of QqTOF instruments over triple quadrupoles is the high mass resolution of TOF, typically around 10,000 ($m/\Delta m$, where Δm is the full peak width at half-maximum (FWHM)). Because electrospray ionization is a continuous process, and TOF analyzer is pulsing ions in time-discrete fashion, some losses of sensitivity occur because of the duty cycle. However, in operational modes requiring the acquisition of full *m/z* range spectra, the sensitivity is more than regained since all *m/z* are recorded in a single TOF spectrum without scanning [31]. Furthermore, in order to maintain the high sensitivity in MS/MS mode, Q3 quadrupole in triple quadrupole machines is usually operated under low resolution settings. However, in QqTOF machines MS/MS spectra are acquired by TOF analyzer, and neither resolution nor sensitivity is compromised in comparison with MS operational mode.

Because of Q0 and a few additional focusing elements, QqTOF mass spectrometers can be equipped with a MALDI source [33] (Figure 3B), or with combined rapidly switchable ESI / MALDI source [34]. This technique is termed

orthogonal MALDI (o-MALDI)-qTOF. In o-MALDI the pulsing ion beam created by high repetition rate laser shots is damped in Q0 and converted into a quasi-continuous beam guided by Q1 and Q2 into an orthogonal TOF instrument in the same manner as described above for electrosprayed ions. Upon passing Q0 ions are moving (relatively) slow and precursor ions can be isolated by Q1 for subsequent MS/MS experiments. o-MALDI instruments enable to combine rapid peptide mass fingerprinting of protein digests with extensive characterization of individual peptide precursors by low-energy collision induced dissociation [35]. Proteins can be identified in all types of sequence databases either by peptide mass mapping or by tandem mass spectra acquired from multiple peptide precursors in the course of a single experiment [36].

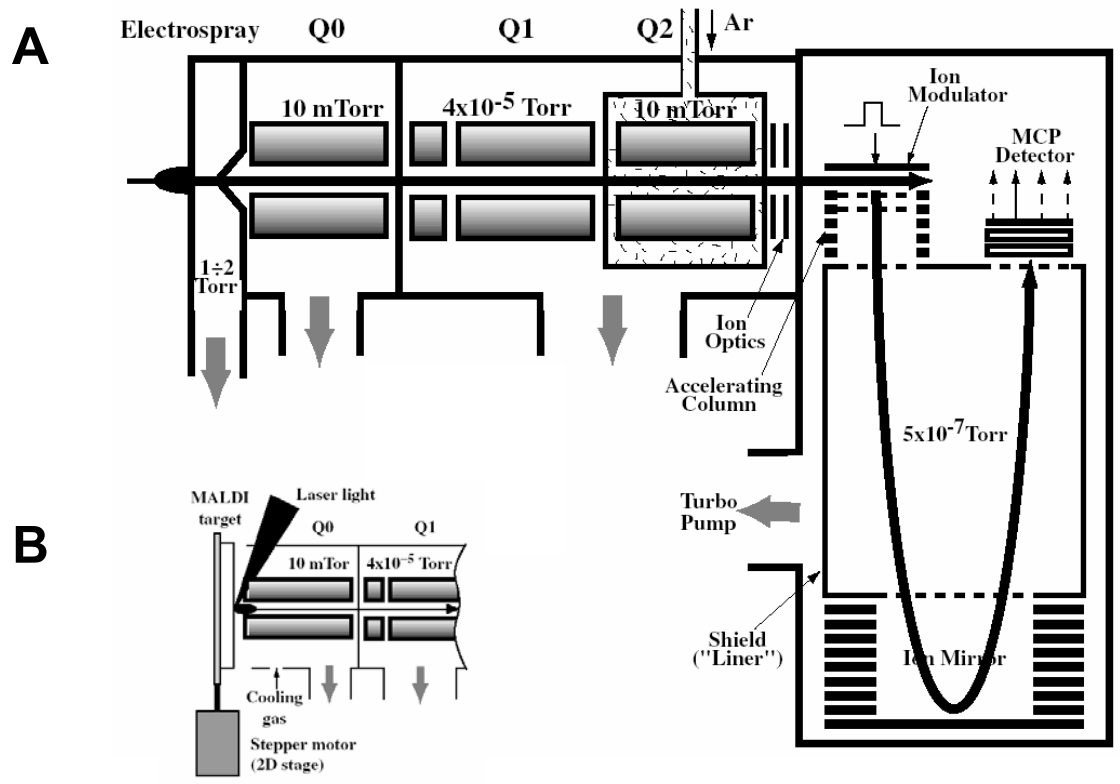


Figure 3. Schematic diagram of the tandem quadrupole time-of-flight mass spectrometer. (A) QqTOF equipped with the electrospray source; (B) QqTOF equipped with the MALDI source (from [31]).

1.1.1.5 Other types of mass spectrometers

In ion trap analysers, ions are first captured or 'trapped' for a certain time interval and then subjected to MS or MS/MS analysis [37]. Ion trap instruments are robust, sensitive, relatively inexpensive and enable rapid highly automated peptide sequencing [38]. A very important feature of ion traps is their ability to perform so called MS^n experiments. Because energy transfer in an ion trap depends on the resonance frequency, only ions of a given *m/z* can be collisionally excited, whereas other ions falling off the resonance frequency (including the fragments of the collisional decomposition of the precursor) remain cool. These ions can, in turn, be collisionally activated, resulting in the new series of fragments. Apparent disadvantage of ion traps rests with their relatively low mass accuracy, typically in the range of 0.2 – 0.4 Da for MS and MS/MS experiments.

Linear or two-dimensional ion trap mass spectrometers have been recently introduced [39, 40]. Ions are stored in a cylindrical volume that is considerably larger than that of conventional three-dimensional traps, thus improving the sensitivity and dynamic range.

Fourier Transform Ion Cyclotron Resonance (FT ICR) instruments are also trapping mass spectrometers, although they capture ions under high vacuum in a high magnetic field. Their strength is the unmatched mass resolution (>100,000), which, if properly used, can result in 0.1 – 1 ppm mass accuracy [41]. But despite of the large potential, high costs, operational complexity and low peptide fragmentation efficiency of FT-MS instruments have limited their routine use in proteomics. Recently hybrid linear ion trap – FT MS instruments have become commercially available. In these hybrid instruments efficient fragmentation is performed in the linear ion trap, whereas FT ICR analyzer is only employed to detect the masses of fragments. Thus, because of high fragmentation efficiency combined with high mass accuracy and high sensitivity these instruments will likely have substantial impact on the entire proteomics field.

MALDI ion source can also be combined with a two-section TOF instrument separated by a fragmentation chamber, a so-called TOF-TOF machine [42]. This instrument offers very fast data acquisition and the sensitivity is not affected by duty

cycle. However, accurate selection of precursor ions and accurate measurement of the masses of fragments are difficult to achieve.

1.1.2 Mass spectrometry based proteomics

1.1.2.1 Analytical strategies

A generic mass spectrometry-based proteomics experiment [19] includes three basic stages: 1) isolation of protein(s) of interest, 2) enzymatic cleavage and 3) mass spectrometric analysis of protein fragments. Despite recent developments in mass spectrometry technology, a "top down" proteomics of intact proteins has a relatively limited scope - because of post-translational processing, masses of intact proteins or of their large fragments are much less specific in the protein identification [43, 44].

Two analytical strategies are typically employed by MS-based proteomics. In the first strategy (Figure 4), proteins are separated by one- or two-dimensional gel electrophoresis, digested *in-gel* by proteolytic enzymes and identified by mass spectrometry (stage 3). To increase the specificity of mass spectrometric detection, peptides could be on-line separated by the method of nanoflow liquid chromatography or capillary electrophoresis.

The limitations of this strategy are dynamic range of detection, variable elution efficiency of digestion products from a polyacrylamide matrix and potential selection against proteins with properties that impede analysis by SDS–PAGE (*e. g*. high or low molecular weight proteins, poorly soluble membrane proteins, extensively modified proteins, etc.[45]).

Multidimensional liquid chromatography integrated on-line with mass spectrometry (MudPIT) provides a powerful alternative to classic two-dimensional electrophoresis [46, 47]. This strategy relies on in-solution digestion of a complex protein mixture, followed by on-line sequencing of even more complex peptide mixture by "fast" mass spectrometers, such as ion traps and quadrupole TOF instruments. This approach circumvents the limitations imposed by two-dimensional gel electrophoresis, and also lends itself to complete automation and scalability. However, when applied to the analysis of protein complexes, this approach lacks important aspects of quantification to determine if novel proteins were present in the stoichiometric or sub-stoichiometric amount, compared with core subunits of the

complex. Also the reproducibility of protein identification in complex mixtures (on average, 70 %) [1, 48] might not allow unambiguous assignment of individual proteins to complexes.

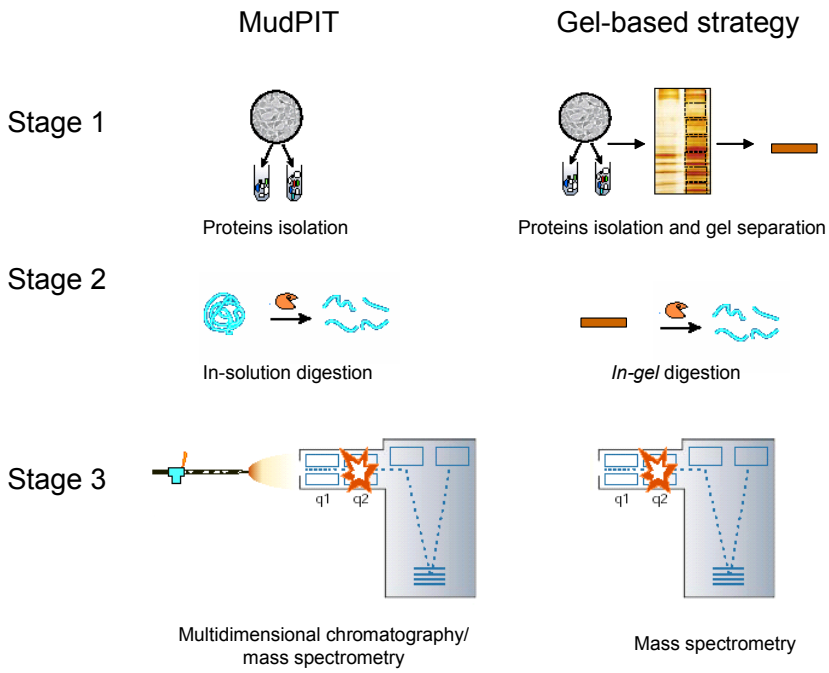


Figure 4. Mass spectrometry based strategies in proteomics (adapted from [19]). A generic mass spectrometry-based proteomics experiment includes three stages: Stage 1 – isolation of proteins of interest; Stage 2 – enzymatic cleavage and Stage 3 – mass spectrometric analysis of protein fragments. MudPIT strategy (left) is based on in-solution digestion of the protein mixture followed by the separation of digestion products by multidimensional liquid chromatography, whereas a gel-based strategy (right) includes SDS PAGE separation of isolated proteins and *in-gel* digestion of individual bands.

1.1.2.2 Protein identification by mass spectrometry

Methods of identification of proteins by mass spectrometry fall into two major categories [49]. First, intact masses of tryptic peptides detected in the unseparated protein digest can be used for probing a database. This approach is termed peptide

mass fingerprinting (PMF or peptide mass mapping) [50] and is applied most efficiently to either individual proteins [51], or simple protein mixtures [52]. Peptide mass fingerprinting is exceptionally powerful when combined with MALDI TOF mass spectrometry, since mass spectra of digests can be acquired rapidly and with high mass accuracy, thus resulting in a very high throughput analysis [53].

Alternatively, MS/MS spectra can be acquired from several precursor ions, and lists of masses of fragment ions together with masses of precursors can be directly submitted to database searching. Search engines, such as Mascot [54] or Sequest [55] correlate the mass list with the mass profile expected for a peptide of a given sequence and then calculate the relative score of this correlation. If several peptides were independently matched to a protein sequence, the confidence of the identification increases. The important feature of these (and similar) algorithms is that searches can be fully automated and do not require manual interpretation of tandem mass spectra. However, in case of poor quality spectra the assignment of fragment ions becomes ambiguous and the actual threshold score, separating true hits from false positives is poorly controlled. Also, in order to identify the protein, exactly the same peptide sequence should be present in a database, and therefore the scope of these methods is mostly limited to species with completely and accurately sequenced genomes.

Tandem mass spectra could also be interpreted and complete or partial peptide sequences deduced [56]. If it is only possible to determine a stretch of sequence consisting of 2-3 amino acid residues, it can be combined with corresponding fragment masses and the mass of the intact peptide into a peptide sequence tag [57]. Searches with peptide sequence tags can be performed in an error-tolerant fashion, so that only parts of the peptide sequence tag match, and the entire process has been recently automated [58].

If the protein sequence is not available in a database, full length sequences of fragmented peptides can be deduced by manual or software-assisted interpretation of tandem mass spectra and used to probe a database by homology searching engines, adapted to utilize peptide sequences produced by mass spectrometry [59, 60].

1.1.3 Quantification by mass spectrometry

1.1.3.1 Relative and absolute quantification approaches

Quantification of proteins by mass spectrometry is typically performed in the relative or absolute fashion.

Relative (or comparative) quantification usually determines relative changes of the amount of a given protein between experimental and control samples. Relative amounts of many proteins can be compared in parallel, thus providing a quantitative overview of the dynamically altered proteome. Proteins can be metabolically labeled with stable isotopes by growing cells in isotopically enriched media. Experimental and control cell pools are then mixed and proteins of interest can be further enriched by sophisticated analytical procedures, which would not affect the accuracy of the analysis, since isotopically labeled and native proteins are likely to behave identically [61, 62]. Alternatively, protein mixtures recovered from experimental and control cells can be separately treated with isotopically labeled and unlabeled chemical probes [63]. The derivatized protein pools are then mixed, digested with enzymes, and modified peptides are enriched by affinity chromatography and quantified by mass spectrometry. In both approaches proteins are quantified by comparing the intensities, or peak areas, of the isotopically labeled and native forms of the same peptide, and the protein amount is averaged if several pairs of peptides from the same proteins were analyzed. Both approaches enable accurate relative comparison of the amount of the same protein, no matters how many different proteins were analyzed in parallel. However, it is impossible to quantify the relative amount of different proteins, since the response of a mass spectrometer strongly depends on the amino acid composition and the sequence of analyzed peptides.

Another analytical approach determines the absolute amount of the analyzed protein, either in moles or in grams per cell or per purification, and therefore the content of different proteins present in the sample can be directly compared. Being more direct, the absolute quantification is also more technically demanding since it is important that peptides with sequences that are almost identical to sequences of peptides from the quantified analyte are employed as internal standards [64, 65]. Peptides that mimic native peptides rendered by the proteolysis of target proteins are chemically synthesized from isotopically-enriched amino acids and are furher

employed as internal standards. These peptides can also be synthetized with covalent modifications (*e.g.,* phosphorylation, methylation, acetylation, etc.) that are chemically identical to naturally occurring posttranslational modifications. The absolute quantification approach (AQUA) reported in [66] uses isotopically labeled synthetic peptides as internal standards to determine quantitatively the absolute amount of proteins and post-translationally modified peptides using single reaction monitoring on a ripple quadrupole mass spectrometer.

1.1.3.2 Peptide labeling with stable isotopes

It is usually not straightforward to correlate the intensity of peptide signals detected by mass spectrometry with the amount of the analyte and therefore internal standards are required. Isotopically enriched peptide standards are produced by the following major methods:

i. Proteins are labeled metabolically by culturing cells in isotopically enriched media (for example, containing ^{15}N ammonia salts, or ^{13}C-labelled amino acids [62, 67]).

ii. Proteins are labeled at specific sites with isotopically encoded reagents. The reagents can also contain affinity tags, allowing for the selective isolation of the labeled peptides after protein digestion. The use of chemistries of different specificity enables selective tagging of classes of proteins containing specific functional groups. Isotope-tagging chemistries specific for sulphydryl groups [63, 68], amino groups [69], the active sites for serine [70] and cysteine hydrolases [71], for phosphate ester groups [72] and for *N*-linked carbohydrates [73] were reported.

iii. Proteins are isotopically tagged by means of enzyme-catalyzed incorporation of ^{18}O atoms from ^{18}O-water during proteolysis. Each peptide produced by the enzymatic cleavage of proteins carried out in ^{18}O-water is labeled at the carboxy-terminus [74, 75].

In each case, labeled proteins or peptides are spiked into solutions of quantified proteins or peptides and their relative abundance is determined. The sensitivity requirements for these approaches are somewhat higher than for straightforward identification because the ion statistics should enable reliable measurements of the ratios of isotopic peaks. The mass difference of peptide pairs generated by metabolic labeling is dependent on the amino acid composition of the

peptide and is therefore variable. The mass difference generated by chemical tagging is one or multiple times the mass difference encoded in the reagent used.

1.2 Proteomics approaches for protein-protein interactions analysis

In the living cell proteins rarely work as isolated entities. They almost always interact with other biomolecules to execute their functions. Networks of such biomolecular interactions constitute the basis of life, and those occurring between proteins play extremely important roles [76]. Therefore the first question usually asked about a new protein - apart from where it is expressed - is to what proteins does it bind? In the past few years extensive efforts were dedicated to charting protein–protein interactions on a genome-wide scale. Two major approaches have been used: yeast two-hybrid screens [77], which detects binary interactions *in vivo*, and biochemical co-purification of complexes using affinity tags, coupled with protein identification by mass spectrometry, which defines protein complexes that comprise a particular protein bait. Fluorescent-based interaction assays were also developed, but have not yet been employed in high-throughput screens.

1.2.1 Affinity purification of protein assemblies

1.2.1.1 Generic strategy, advantages and limitations

The protein itself can be used as a bait to isolate its binding partners. Such protein interaction experiments comprise three essential components: bait presentation, affinity purification of the complex, and analysis of the bound proteins, *e.g.* by mass spectrometry. Ideally, endogenous proteins can serve as baits if antibodies or other reagents are available that allow the specific isolation of the protein along with its interaction partners. Unfortunately, collections of antibodies are not comprehensive and many available antibodies lack the affinity and/or specificity to enable efficient immunoprecipitation of corresponding baits from whole cell lysates.A more generic strategy is to ‘tag’ the proteins of interest. Candidate genes can be tagged with sequences coding for epitopes that are recognized by monoclonal antibodies. The epitope-tagged protein, together with associated proteins, is recovered by the immunoprecipitation from a whole cell lysate. Proteins that interact with the

tagged subunit are separated by gel electrophoresis and identified by mass spectrometry [78].

Compared with two-hybrid and chip-based approaches, this strategy has the advantages that the fully processed and modified protein can serve as a bait, that the bait and preys interact in the native environment and cellular location, and that multicomponent complexes can be isolated and analyzed in a single analysis. However, low affinity, transient interactions and those dependent on the specific cellular environment might be lost in affinity purification experiments. Bioinformatics methods [79], correlation of data with those obtained by other methods [80], or iterative mass spectrometric measurements possibly complemented by chemical crosslinking [81] of subunits can help to further elucidate direct interactions and overall topology of multiprotein complexes.

Affinity purification methods were successfully implemented to study protein-protein interactions on a genome scale in the budding yeast [1, 2] and in many smaller scale projects in a variety of organisms.

1.2.1.2 Commonly used affinity tags

What tagging system is required for the efficient isolation of protein complexes? The tag should be non-toxic, easy to introduce, should keep expression of the bait protein at the natural level and should not affect its function. Protein purification should be performed under mild conditions, which enable to preserve weak protein-protein interactions, and should be reproducible, rapid and cost efficient.

Expression of the tagged protein at close to physiological levels can be achieved in a limited number of species, most notably *S. cerevisiae*, replacing the endogenous gene in the genome with a gene coding the tagged protein by homologous recombination. In mammalian cells, in which the expression of tagged proteins from native promoters is difficult, proteins are usually expressed after transient transfection or in stable cell lines generated by conventional selection. Transient or stable transfections usually result in tagged protein expression levels that are different from the expression of untagged, endogenous proteins. These expression systems are prone to artifacts prompted by non-physiological levels of bait proteins, such as association with heat shock proteins, proteasome, etc.. Considerable efforts have been devoted to

developing tagging systems optimized for the analysis of protein complexes (reviewed in [82]). Two major types of tags have been described:

1. Tags supporting single-step purifications. These tags are convenient to use and isolations typically render higher yields of baits. Glutathione S-transferase (GST) [83] binds strongly to glutathione-agarose and is eluted with glutathione, allowing very high levels of purification in one step; however, it is known to dimerize and this might perturb the function of the bait protein.

His-tag (incorporated hexahistidine sequence) binds to immobilized nickel affinity columns and is eluted with imidazole [84, 85]. This tag is very simple and small in size, but it results in relatively modest purification yield, compared to other tags.

Calmodulin-binding peptide binds to calmodulin-agarose columns and is eluted with EGTA. Purification is efficient, but the structure of proteins binding divalent metal ions may be affected. Other commonly used single-step tags against which good quality antibodies are commercially available include haemagglutinin (HA-tag), Myc and FLAG –tags [86] and Ig-binding domain of protein A [87].

2. Tags supporting two sequential affinity purification steps (tandem affinity purification) combine two different tags on the same protein, which are spaced by an enzyme-cleavable linker sequence [88, 89]. These tags significantly reduce background, but probably result in the loss of transient and weakly binding partners during the purification procedure.

1.2.1.3 The TAP method

The TAP (tandem affinity purification) method is one of the most promising approaches for isolation of protein complexes. The TAP double tag [89] consists of the calmodulin binding peptide (CBP) and two IgG binding domains of protein A spaced by a cleavage site of tobacco etch virus (TEV) proteinase (Figure 5A).

These two components were selected upon semi-quantitative evaluation of seven high-affinity tags: FLAG, a tag with two IgG-binding units of protein A of *Staphylococcus aureus* (ProtA), Strep tag, His tag, calmodulin-binding peptide (CBP), and chitin-binding domain (CBD) [148, 149]. None of these tags impaired the protein function, however only ProtA and CBP tags allowed efficient recovery (roughly 80%

and 50%, respectively) of the fusion protein, which was present at the low-picomole level in a complex mixture of proteins [89].

CBP tag allowed efficient selection and specific release of bound proteins from the affinity column under mild experimental conditions. In contrast, ProtA can only be released from matrix-bound IgG under denaturing conditions at low pH. Therefore, specific TEV protease recognition sequence was inserted upstream of ProtA. TEV cleavage can be performed under native non-denaturing conditions leaving a part of the tag bound onto the beads after the first step of purification.

The TAP method (Figure 5B) involves the fusion of the TAP tag to the gene encoding for the target protein followed by the incorporation of the construct into the host cells or organism. The fusion protein and associated components are recovered from whole cell lysates by the affinity purification on IgG matrix. After washing, TEV protease is added to release the bound material. The eluate is incubated with calmodulin-coated beads in the presence of calcium. This second affinity step is required to remove the TEV protease as well as contaminants remaining after the first affinity purification. After washing, the bound material is released from the beads by EGTA. Purified proteins are concentrated and separated by SDS-PAGE.

The TAP tag can be fused to the C-terminus or N-terminus of the bait protein, although the latter method is more technically demanding [120].

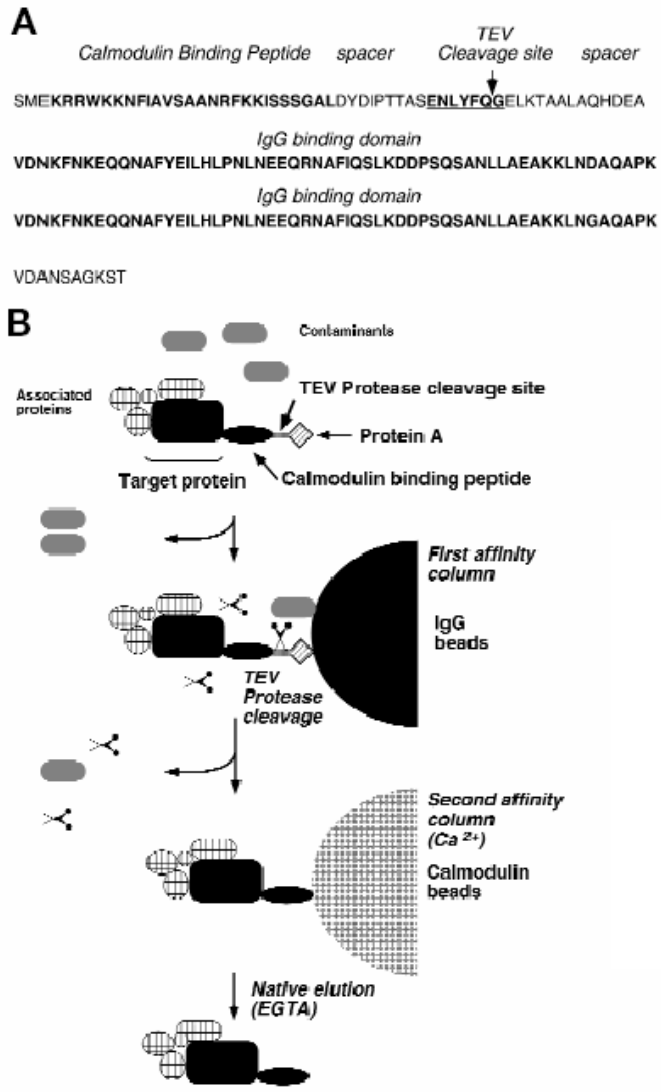


Figure 5. The TAP purification method. (A). Sequence and structure of the TAP tag. The various domains constituting the TAP tag are indicated. (B). Overview of the TAP procedure (from [89]).

1.2.2 Two hybrid approach: generic strategy, advantages and limitations

Two-hybrid assay first introduced by Fields and Song [77] provides a genetic approach to the identification and analysis of protein–protein interactions. It relies on the modular nature of many eukaryotic transcription factors, which contain both a site-specific DNA-binding domain and a transcriptional-activation domain that recruits the transcriptional machinery. In this assay, hybrid proteins are generated that fuse a protein X to the DNA-binding domain and protein Y to the activation domain

of a transcription factor. Interaction between X and Y reconstitutes the activity of the transcription factor and leads to expression of reporter genes with recognition sites for the DNA-binding domain (Figure 6A). In the typical practice of this method, a protein of interest fused to the DNA-binding domain (the so-called 'bait') is screened against a library of activation-domain hybrids ('preys') to select interacting partners (Figure 6B).

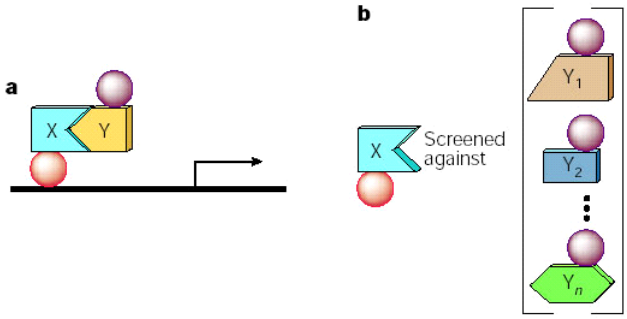


Figure 6. Principal scheme of two-hybrid approach. (adapted from [90]). Red sphere – DNA-binding domain, lilac sphere - transcriptional-activation domain, 'X' – bait protein, 'Y' – prey proteins. (A). Interaction between bait and prey proteins reconstitutes activity of transcription factor and leads to expression of reporter genes. (B). Bait protein is normally screened against a library of potential preys.

Key advantages of the two-hybrid assay are its sensitivity and flexibility. Sensitivity leads to the detection of interactions with dissociation constants around 10^{-7} M [90], in the range of most weak protein interactions found in the cell, and is more sensitive than co-purification, which requires stability of a complex through dilution from cell lysis, and through subsequent purification steps. This sensitivity also allows detection of certain transient interactions.

Being sensitive and flexible, the method, however, has a number of limitations. The two-hybrid approach cannot detect interactions requiring three or more proteins and those depending on posttranslational modifications. It is not suitable for the detection of interactions involving membrane or secreted proteins, proteins that activate transcription when fused to a DNA-binding domain or that fail to fold correctly [90]. Thus, due to these limitations genome-wide two-hybrid projects may miss most (as much as 90%) of known interactions [91]. Another major concern of two-hybrid approach is false positives, which result from spurious transcription that

occures independently from bait-prey interaction. Therefore, detecting Y2H interactions does not automatically guarantee their physiological relevance. Several analyses of genome-wide two-hybrid screens suggest that only about 50% of discovered interactions are *bona fide* [79, 90].

Despite of these limitations, genome-wide two-hybrid studies have been successfully carried out in *S. cerevisiae* [92-94], *Helicobacter pylori* [95] and *C. elegans* [96].

1.2.3 Fluorescence resonance energy transfer (FRET) method

FRET is a nonradiative process whereby energy from an excited donor fluorophore is transferred to an acceptor fluorophore that is located within 60 Å of the excited fluorophore [97]. After excitation of the first fluorophore, FRET is detected either by emission from the second fluorophore using appropriate filters, or by alteration of the fluorescence lifetime of the donor. Two fluorophores that are commonly used are variants of green fluorescent protein (GFP): cyan fluorescent protein (CFP) and yellow fluorescent protein (YFP). A number of protein interactions have been demonstrated in cells by FRET microscopy [98-101].

The advantage of the method is that measurements can be carried out in living cells and can potentially detect transient interactions and post-translation modifications, however it has not been yet adopted for genome-wide screens.

1.2.4 Other methods

Two-hybrid and affinity isolations techniques are focused on the detection of physical binding between proteins, whereas other methods seek to predict functional associations, e.g. between a transcriptional regulator and the pathway it controls. In many cases, such functional associations do take the form of physical binding.

For example, two nonessential genes that cause lethality when mutated at the same time form a synthetic lethal interaction. Such genes are often functionally associated and their encoded proteins may also interact physically. This type of genetic interaction is currently being studied in an all-versus-all approach in the yeast [102].

Using bioinformatics, available genomes can be screened for different types of interaction evidence *in silico*. For example, in prokaryotic genomes interacting proteins are often encoded by conserved operons [103, 104], or interacting proteins have a tendency to be either present or absent together in fully sequenced genomes [105], that is, to have a similar 'phylogenetic profile'. Also seemingly unrelated proteins may sometimes be found fused into one polypeptide chain, indicating their possible physical interaction [106]. Bioinformatic analysis is fast, inexpensive and expands the coverage as more genomes are sequenced. However, it requires a framework for assigning orthology between proteins, failing where orthology relationships are not clear; and so far it has focused mainly on prokaryotes.

The most effective and reliable approach for studying protein-protein interactions is a combination of different methods, such as affinity purification and mass spectrometry with two-hybrid screens [80] followed by bioinformatic cross-correlation of independently derived datasets [79].

1.2.5 Databases of protein-protein interactions

Protein–protein interaction databases employ various approaches to represent information from the different experimental techniques and varying resolutions of spatial and temporal behavior for protein complexes. Databases, such as DIP [107] and BIND [108] store many details about the experiments, and LiveDIP [109] can even model most details of biological complexes, if sufficient experimental data are available. Simplified and less redundant databases, such as YPD and PombePD [110] and the MIPS collection of protein complexes [111, 112] are popular, because they underwent manual curation.

1.3 Aim of the work

The major goal of this study was to develop an experimental proteomics approach for deciphering protein complexes and protein interaction networks in the budding and fission yeasts.

TAP isolation of protein assemblies followed by SDS PAGE separation and mass spectrometric identification of interactors was chosen as a basic proteomics startegy. Therefore key elements of the proteomic routine, such as gel storage, visualization of protein bands, mass spectrometric analysis, as well as analytical aspects of the TAP purification method were studied and optimized to improve the sensitivity and reliability of the identification of protein interactors.

Once developed, the strategy was applied for the comparative characterization of protein complexes and segments of protein interaction networks in the budding and fission yeasts.

2. Materials and Methods

2.1 Materials

2.1.1 Mass spectrometers

MALDI TOF REFLEX IV mass spectrometer (Bruker Daltonik GmbH, Germany) equipped with Scout 384 ion source and AnchorChip 600/384 target (Bruker Daltonik GmbH).

Modified Q STAR Pulsar *i* hybrid quadrupole time-of-flight mass spectrometer (MDS Sciex, Toronto, Canada) equipped with a nanoelectrospray ion source (Proxeon Biosystems A/S, Odense, Denmark). In our instrument the exit grid placed between the collision cell and the entrance slit into the TOF chamber has been replaced by a short DC quadrupole. This increased dramatically the sensitivity of the instrument and the stability of ion beam focusing.

Prototype MALDI QqTOF mass spectrometer [35] built in the laboratory of Prof. Kenneth Standing in the University of Manitoba, Winnipeg, Canada. The instrument consisted of a specially designed MALDI ion source [33] mounted onto a prototype hybrid quadrupole time-of-flight mass spectrometer. Laser pulses were generated by a nitrogen laser model 337 ND with energy per pulse of 250 μJ. A 0.2 mm core diameter fused-silica optic fiber delivered laser pulses to the target. The laser beam was focused at the target into a spot about 0.3 mm x 1 mm. A 10 kV acceleration voltage was used in the TOF part of the instrument. The orthogonal injection pulse repetition rate was set to 8.5 kHz.

API III triple quadrupole mass spectrometer (MDS Sciex, Toronto, Canada) equipped with a nanoES ion source [29] designed in EMBL Heidelberg, Germany)

2.1.2 Supporting equipment

Micropipette puller, model P-97 (Sutter Instrument Co., USA)
Sputter coater SC7620 (Polaron)
Benchtop mini-centrifuge PicoFuge (Stratagene)
NanoES minipurification holder (Protana, Odense, Denmark)
Microscope Zoom2000 (Leica, Germany)

2.1.3 Gel electrophoresis equipment

Mini-Protean II gel running system (7 cm x 10 cm minigels) (BioRad)
Protean II xi Cell gel running system (BioRad)
Multicasting module for minigels (BioRad)
GelAirDryer (BioRad)
Vacuum filtering system Stericup (Millipore)

2.1.4 Materials

Acrylamide/bisacrylamide	BioRad
Acetic acid	Merck
Acetone	Sigma
Acetonilrile, HPLC grade	Merck
Ammonium persulfate (APS)	Sigma
Ammonium bicarbonate	Sigma
Bovine serum albumin (BSA)	Sigma
Broad range protein MW standard	BioRad
Borosilicate glass capillaries	Clark Instruments, UK
Calcium chloride	Sigma
Cellophane for gel drying	BioRad
Coomassie Brilliant Blue R250	Serva Electrophoresis GmbH
Cytochrome c	Sigma
Dithiotreitol (DTT)	Sigma
2,5-Dihydroxybenzoic acid (DHB)	Sigma
Electrophoresis running buffer x10	BioRad
Ethanol	Merck
Formic acid	Merck
Formalin (37% formaldehyde)	Merck
GFP ([Glu]-fibrinogen peptide)	Sigma
Glycerol	Merck
Glycine	Merck
HCl	Merck
$H_2{}^{16}O$, HPLC grade	Merck
$H_2{}^{18}O$, analytical grade	Cambridge Isotopic Laboratories
Imidazole	Sigma
Iodoacetoamide	Sigma
Methanol	Merk
MilliQ water	Millipore, Bedford, MA
Myoglobin	Sigma
2-mercaptoethanol	Sigma
Nitrocellulose	BioRad
PCR minitubes, 0.65ml	Roth
Poros R2 sorbent	Perseptive Biosystems, US
Potassium ferricyanide	Sigma
2-propanol	Merck
R-cyano-4-hydroxy-trans-cinnamic acid (CHCA)	Sigma; Bruker
SDS	Sigma
Sodium thiosulfate	Sigma
Sodium carbonate	Sigma
Silver nitrate, SigmaUltra grade	Sigma
TEMED	BioRad
Trifluoroacetic acid (TFA)	Sigma
Tris base	Serva
Trypsin, modified, sequencing grade	Promega, Germany
Trypsin, unmodified, sequencing grade	Roche Diagnostics GmbH
Zn sulfate	Sigma

2.1.5 Buffers and Solutions

Running gel buffer (12%) (two mini gels)	4ml AA/BIS (30%/0.8%) 2.5ml Tris/HCl (1.5M/pH8.8) 100µl SDS (10%w/v) 50µl APS 10µl TEMED 3.2ml H_2O
Stacking gel (4%) (two mini gels)	0.65ml AA/BIS (30%/0.8%) 1.25ml Tris/HCl (0.5M/pH6.8) 100 µl SDS (10%w/v) 60µl APS 10µl TEMED 2.95ml H_2O
Sample loading buffer	1.25ml Tris/HCl (0.5M/pH6.8) 2.5 Glycerol 2.0ml SDS (10%w/v) 0.2ml Bromphenol blue (0.5%w/v) 3.55ml H_2O 50ml 2-Mercaptoethanol
Coomassie Blue staining	0.2% Coomassie Brilliant Blue R250 (w/v) 45% methanol (v/v) 10% acetic acid (v/v) 45% water (v/v)
Coomassie Blue destaining I	45% methanol (v/v) 10% acetic acid (v/v) 45% water (v/v)
Coomassie Blue destaining II	50% acetonitrile (v/v) 0.1M ammonium bicarbonate 50% water (v/v)
Gel fixing solution	45% methanol (v/v) 10% acetic acid (v/v) 45% water (v/v)
Silver staining developing solution	0.04% formalin (w/v) 2% sodium carbonate (w/v) in water
In-gel reducing buffer	10mM DTT 100mM ammonium bicarbonate in water
In-gel alkylation buffer	55mM iodoacetamide 100mM ammonium bicarbonate in water
Trypsin digestion buffer	50mM ammonium bicarbonate 5mM calcium chloride 12.5ng/µl Trypsin

2.2 Methods

2.2.1 Protocols for staining of SDS PAGE gels

2.2.1.1 Coomassie staining

After electrophoresis gels were placed in 0.2% Coomassie Brilliant Blue R250 in 50% methanol in water containing 5% acetic acid for 1 h (for minigels) or overnight and destained for several hours with the same solvent, excluding the dye. For complete removal of Coomassie prior to silver staining gels were destained with 50% acetonitrile solution in water containing 0.1M ammonium bicarbonate.

2.2.1.2 Silver staining [23]

After electrophoresis, the gel slab was fixed in 50% methanol, 5% acetic acid in water for 20 min. It was then washed for 10 min with 50% methanol in water and additionally for 10 min with water to remove the remaining acid. The gel was sensitized by incubating for 1 min in 0.02% sodium thiosulfate then rinsed with two changes of water for 1 min each. After rinsing, the gel was submerged in chilled 0.1% silver nitrate solution and incubated for 20 min at 4 °C. After incubation, silver nitrate solution was discarded and the gel slab was rinsed twice with water for 1 min and then developed in 0.04% formalin in 2% sodium carbonate with intensive shaking. After the developing solution turned yellow, it was discarded and replaced with a fresh portion. After the desired intensity of staining was achieved, development of the gel was terminated by discarding the reagent, followed by washing of the gel slab with 5% acetic acid. Silver-stained gels were stored in 1% acetic acid at 4 °C until analyzed.

2.2.1.3 Destaining of silver stained gels

Protein bands were excised from the gel and destained with potassium ferricyanide and sodium thiosulfate as described [113]. Gel pieces were rinsed with 25mM ammonium bicarbonate, dehydrated by acetonitrile and dried down in a vacuum centrifuge.

2.2.1.4 Negative (Zn-Imidazole) staining [114]

After electrophoresis, gels were rinsed with water twice (1min) and placed in 0.2M imidazole solution (1.4g imidazole to 100ml H_2O) for 10min. After a quick

rinse with water gels were developed by 0.05M Zn sulfate solution (1.44g $ZnSO_4x7H_2O$ in 100ml H_2O).Staining was stopped by intensive washing with water. A running buffer (without SDS) or 5% formic acid can destain negatively stained bands.

2.2.1.5 Gel archiving

After staining, gels were rinsed with water several times, incubated in Gel Drying Solution (BioRad) for 1h, placed between two cellophane sheets (air bubbles should be carefully removed!) and dried in GelAirDryer (BioRad) for 2 hours. Dried gels were stored at room temperature in folders. For easy removal of cellophane before *in-gel* digestion, protein bands from dried gels were excised and incubated in water for 20 min.

2.2.2 *In-gel* sample preparation

2.2.2.1 *In-gel* reduction and alkylation of proteins

After a band (spot) was excised and cut into pieces of approximately 1mm x 1mm x 1mm they were dehydrated by adding neat acetonitrile. Acetonitrile was then aspirated and gel pieces were dried in a vacuum centrifuge. A volume of 10 mM dithiotreitol (DTT) in 100 mM NH_4HCO_3 sufficient to cover gel pieces was added, and proteins were reduced for 1 hour at 56 °C. After chilling down to room temperature, DTT solution was replaced with roughly the same volume of 55mM iodoacetamide in 100 mM NH_4HCO_3. After 45 min incubation at ambient temperature in the dark with occasional vortexing, gel pieces were washed with 50-100 μL of 100 mM NH_4HCO_3 for 10 min, dehydrated by adding neat acetonitrile, rehydrated in 100 mM NH_4HCO_3, and shrunk again by adding acetonitrile. The liquid phase was aspirated and gel pieces dried in a vacuum centrifuge.

2.2.2.2 *In-gel* digestion: conventional protocol [23]

Gel pieces were re-hydrated in a digestion buffer containing 50 mM NH_4HCO_3, 5 mM $CaCl_2$, and 12.5 ng/μL of unmodified trypsin (Roche Diagnostics, Germany) in ice bracket. After 45 min, the supernatant was removed and replaced with 5-10 μL of the same buffer, but without trypsin, to keep the gel pieces wet during enzymic cleavage (37 °C, overnight).

2.2.2.3 *In-gel* digestion: accelerated protocol [115]

Proteins were digested *in-gel* with the modified trypsin (Promega), without prior destaining, reduction, and alkylation. Gel pieces were washed with water for 5 min and dehydrated in neat acetonitrile for 20 min. Acetonitrile was aspirated, and gel pieces were dried in a vacuum centrifuge and re-hydrated for 60 min in 1.5 μM solution of the modified trypsin in 50 mM ammonium bicarbonate buffer at 4 °C. The excess of the enzyme solution was aspirated, and the digestion was carried out for 30 min at 58 °C. The reaction was stopped by adding 1 μL of 5% formic acid.

2.2.2.4 Peptide extraction after *in-gel* digest

Peptides were extracted by one change of 20 mM NH_4HCO_3 and three changes of 5% formic acid in 50% acetonitrile (20 min for each change) at room temperature, all fractions were pooled together and dried down in a vacuum centrifuge.

2.2.3 Sample preparation for mass spectrometry

2.2.3.1 Preparation of MALDI probes by fast evaporation method [22]

Matrix (α-cyano-4-hydroxy-*trans*-cinnamic acid) and nitrocellulose were dissolved in acetone/2-propanol (1:1 v/v) in concentration of approximately 20 and 5 g/L, respectively. Approximately 0.5 μL of the matrix/nitrocellulose solution was deposited on a polished stainless steel target, where it spread rapidly, allowing fast evaporation of the solvent. An aliquot of 0.5 μL of analyte solution was deposited onto these matrix surfaces, and the solvent was allowed to dry at ambient temperature. Upon drying probes were rinsed by placing a 5-10 μL volume of water onto the matrix surface. The liquid was left on the sample for 10 s and was then blown off by pressurized air. Washing procedure was repeated twice.

2.2.3.2 MALDI probe preparation on AnchorChip™ targets

An aliquote of 0.6 μL withdrawn from an *in-gel* digest and 0.6 μL of matrix solution (2 mg/mL 1-cyano-4-hydroxycinnamic acid in 2.5% trifluoroacetic acid/acetonitrile, 1:2 v/v) were mixed at the surface of an AnchorChip 600/384 target

(Bruker Daltonik GmbH), allowed to dry down at ambient temperature and subsequently washed with 5% formic acid.

2.2.3.3 Probe preparation for MALDI QqTOF analysis [36]

Dried mixture of tryptic peptide was re-dissolved in 5μL of 5% formic acid. A matrix solution of 160 mg/mL 2,5-dihydroxybenzoic acid (DHB; Sigma) was prepared in a 1:3 mixture of acetonitrile and water. A total of 0.6 μL of the matrix solution was first deposited on the target to form a spot 2 - 3 mm in diameter. Samples were dissolved in 2-5 μL of 5% formic acid, and typically, a 0.3 - 0.6 μL aliquot was deposited on top of the matrix spot. No further washing of DHB crystals was performed.

2.2.3.4 Sample preparation for the analysis by nanoES MS/MS [26]

Gel pieces were extracted as described in section 2.2.2.4 and the combined extracts were dried down in a vacuum centrifuge and re-dissolved in 10 μL of 5 % formic acid. 5 μL of POROS R2 slurry prepared in methanol into the pulled glass capillary (here and further down referred as a "column"). Beads were spun down and then the pulled end of the column was opened by gentle touching against a bench top. The beads were washed with 5 μL of 5% formic acid and additionally 5 μL of formic acid was passed through the capillary to test if the liquid can easily be spinned out of the column by gentle centrifuging. Otherwise, the column end was opened wider. The column was mounted into the micropurification holder and the sample was passed through the column by centrifuging. Adsorbed peptides were washed with another 5 μL of 5% formic acid. The column and the nanoelectrospray needle were aligned in the micropurification holder and peptides were eluted directly into the needle with 1 μL of 60% of methanol in 5% formic acid by gentle centrifuging. The spraying needle with the sample was mounted into the nanoelectrospray ion source and mass spectra were acquired.

2.2.4 Data acquisition and processing

2.2.4.1 Data acquisition and processing software

LaserOne (EMBL, Heidelberg, Germany).

BioMultiView 1.4 (Sciex, Canada)

These two programs were operated on the Apple Macintosh platform

Analyst QS SP6 (MDS Sciex, Concord, Canada)

XMmass 5.1.1 version (Bruker Daltonk GmbH, Germany)

BioTools 2.1 version (Bruker Daltonk GmbH, Germany)

These programs were operated on the Windows platform.

2.2.4.2 Data acquisition and processing on REFLEX IV MALDI TOF mass spectrometer

All spectra were acquired under operator control. Accelerating voltage was set at 16.8kV, reflector voltage 23kV, pulsed ion extraction was 400 ns. Laser fluence was adjusted by the operator to avoid saturation of the detector. Number of shots per sample varied depending on the amount of peptide material. Each MALDI spectrum of a protein digest was calibrated internally using several matrix and trypsin autolysis peaks as references [116]. Spectra were smoothed (3 data points Savitzki-Golay filter), monoisotopic peaks of peptide ions were selected manually, masses of known trypsin autolysis peptide and typical contaminant masses were subtracted from the peak lists.

2.2.4.3 Data acquisition and processing on a prototype MALDI QqTOF mass spectrometer

MS-only TOF spectra were acquired at thelaser repetition rate of 5-10 Hz for time intervals typically less than 1 min with no adjustment of the laser fluence. Tandem mass spectra were acquired at 10-20 Hz, also without adjustments in the laser fluence. Spectra acquisition time varied depending on the ion current, but rarely exceeded 5 min. The width of the mass window for Q1 was set to ca 2 Da for m/z 500, increasing to 4 Da at m/z 3000. In all experiments, argon was used as a “cooling” gas in Q0 and as a collision gas. The collision energy was set by applying the accelerating voltage at the entrance of the collision cell by the rule 0.6 V/Da and then adjusted manually to obtain a desirable fragmentation pattern. The instrument was calibrated

externally and no post-acquisition recalibration of MS and MS/MS spectra was performed. *PredictSequence* program (a part of the BioMultiview 1.4 software, MDS Sciex) was applied for the interpretation of MS/MS spectra.

2.2.4.4 Data acquisition and processing on API III triple quadrupole mass spectrometer

After desalting and concentration of the sample as described in section 2.2.3.4 spraying was initiated by applying 600 – 800 V to the spraying needle and Q1 spectrum of the peptide mixture was acquired. Q1 scans were performed with a 0.1 Da mass step. Collision gas was then turned on and the spectrum in the precursor scan mode (scanning for precursor ions producing fragment ions with *m/z* 86 upon their collisional fragmentation) was acquired [27]. Spraying was stopped by dropping spraying voltage and air pressure applied to the spraying capillary. The spraying capillary was moved away from the inlet of the mass spectrometer. The acquired spectra were compared with the spectra acquired from the control sample and precursor ions for subsequent tandem mass spectrometric sequencing were selected. Spraying was then reestablished and tandem mass spectra from selected precursor ions were acquired. For operation in the MS/MS mode, Q1 was set to transmit a mass window of 2 Da, and spectra were accumulated with 0.2 Da mass steps. Resolution was set so that fragment masses could be assigned with the mass accuracy better than 0.5 Da. The collision energy was tuned individually for each peptide to obtain the best possible MS/MS spectra as described in [28]. The instrument was calibrated externally according to the instructions from the manufacture and no post-acquisition recalibration of MS and MS/MS spectra was performed.

2.2.4.5 Data acquisition on a QSTAR Pulsar *i* quadrupole time-of-flight mass spectrometer equipped with a nanoES ion source.

Nanoelectrospray analysis of *in-gel* digests was performed as described for a triple quadrupole instrument (1.2.4.4), however no precursor ion scanning was applied. High resolution of the TOF analyzer enabled to distinguish multiply charged isotopically resolved peptide ions among broad and irregular-shaped peaks originating from chemical background [117]. Collision energy was tuned individually for each peptide to obtain the most informative pattern of fragment ions. The instrument was calibrated in MS/MS mode using a synthetic peptide prior each experiment.

Uninterpreted spectra were converted into MASCOT generic format (*.mgf) using a script from the Analyst QS software. On several occasions spectra were interpreted manually and error-tolerant or sequence similarity searches [118] were applied.

2.2.5 Database searching software and settings

2.2.5.1 MASCOT software

MASCOT v. 1.8 software [54] installed at local server was used for database searching with both MALDI peptide mass maps and uninterpreted tandem mass spectra. For searches with tandem mass spectra mass tolerance was set at 2Da for masses of peptide precursors and at 0.05 Da for masses of fragment ions. For searches with peptide mass fingerprints mass tolerance was set at 100 ppm. In both type of searches a conventional set of variable and fixed modifications of amino acid residues was applied. Searches were performed against a non-redundant database (MSDB), which was downloaded from European Bioinformatics Institute (EBI) (ftp://ftp.ncbi.nlm.nih.gov/blast/db/).

2.2.5.2 PeptideSearch

PeptideSearch v. 3.0 software developed in EMBL was used for searches with MALDI peptide mass maps and error-tolerant searching with partially interpreted tandem mass spectra (peptide sequence tags [57]). For database searches with sequence tags deduced by manual interpretation of tandem mass spectra acquired on a triple quadrupole machine, mass tolerance was set at 1 Da for masses of peptide precursors and at 0.5 Da for the masses of fragment ions. No limitations on protein molecular weights and calculated *pI* were applied. A second pass searching routine [52] was applied to match modified peptides in MALDI peptide mass maps.

2.2.5.3 MS BLAST

MS BLAST sequence similarity searches were performed by WU-BLAST2 program (Gish, W. (1996-1999) http://blast.wustl.edu) at the EMBL server: http://dove.embl-heidelberg.de/Blast2/. The following settings were applied [119]: Program: blast2p; Database: nrdb95; Matrix: PAM30MS; “Expect”: 1000; “Other advanced options”: -nogap -hspmax 100 -sort_by_totalscore -span1. Search was performed against a comprehensive non-redundant database nr95 .

2.2.5.4 MS-Tag

Database searching using mass lists deduced from tandem mass spectra acquired on the MALDI QqTOF instrument was performed by MS-Tag program as described [36]. MS-Tag program is a part of ProteinProspector (UCSF Mass Spectrometry facility) and is accessible at http://prospector.ucsf.edu.

2.2.6 Quantification experiments

2.2.6.1 Purification of $H_2{}^{18}O$

Chemical purity of commercially available $H_2{}^{18}O$ is lower than 95% w/w and it is unsuitable for the quantification because of heavy chemical noise observed in mass spectra. 0.5 mL portions of $H_2{}^{18}O$ were purified by microdistillation in a sealed glass apparatus as described [26] and stored at –20°C in 15 µL aliquots until used. Each aliquot was used only once.

2.2.6.2 Preparation of the standard

To prepare a standard mixture of ^{18}O-labeled peptides, a solution of 0.14 pmol/ µL BSA in 25 mM ammonium bicarbonate buffer in $H_2{}^{18}O$ was digested overnight at 37°, enzyme : substrate ratio 1:10 (w/w).

2.2.6.3 Sample preparation for the quantification by MALDI TOF

Protein bands excised from one polyacrylamide gel were in parallel *in-gel* digested with trypsin as described in 2.2.2.2. Gel pieces were extracted with 5% formic acid and acetonitrile and the extracts were dried down in a vacuum concentrator. Tryptic peptides were re-dissolved in 10 µL of 10 % formic acid. 2 µL aliquot was withdrawn and mixed with 1 µL of ^{18}O-labeled mixture of peptides (internal standard) prepared as described in 2.2.6.2. Four 0.5 µL aliquots of every mixed sample were analyzed in parallel by MALDI MS as described in 2.2.4.2. The determined relative concentrations of peptides were averaged. The relative standard deviation of the concentrations in all series of measurements was better than 20 %.

2.2.6.4 Deconvolution of ^{18}O-labeled profiles

In several cases ^{18}O-labeled profiles were deconvoluted as described by Havlis *et al* [115]. MS-Isotope program (a part of ProteinProspector UCSF Mass

Spectrometry facility), accessible at http://prospector.ucsf.edu, was employed for calculating the ratios of intensities of isotopic peaks in MALDI spectra of native peptides.

2.2.7 Analysis of TAP-purified protein mixtures

2.2.7.1 TAP purification routine

Protein TAP-tagging in *S. cerevisiae* and *S. pombe* was performed in collaborating laboratories in EMBL (Group Leaders Drs. B. Seraphin and A.F. Stewart) and in Technische Universität Dresden – BIOTEC (Prof. A.F. Stewart) as described [120, 121]. The original TAP tagging method was applied with minor adjustments in *S.pombe* as described [122, 123]. Mixtures of co-purified proteins were separated by electrophoresis using gradient (6-18 %) one-dimensional polyacrylamide gels and visualized by staining with Coomassie.

2.2.7.2 *In-gel* digestion of TAP-purified proteins

All visible protein bands were excised from SDS PAGE gels and proteins *in-gel* digested as described in 2.2.2.2 and 2.2.2.3. Reduction and alkylation steps were omitted.

2.2.7.3 Identification of TAP-purified proteins

Gel separated proteins were identified by MALDI TOF peptide mapping and nanoelectrospray tandem mass spectrometric sequencing combined in a layered approach [124]. 1µL aliquots withdrawn from their *in-gel* digests were analyzed by MALDI TOF peptide mass mapping. If no conclusive identification was achieved, gel pieces were extracted with 5% formic acid and acetonitrile. Unseparated mixtures of recovered tryptic peptides were sequenced by nanoelectrospray tandem mass spectrometry on the QSTAR Pulsar *i* quadrupole time - of -flight mass spectrometer or on the API III triple-quadrupole mass spectrometer.

Database searches with MALDI TOF peptide mass maps and with uninterpreted tandem mass spectra were performed against a database of *S.pombe or S.cerevisiae* proteins using *Mascot* software. Hits with the MOWSE score exceeding the threshold scores suggested by *Mascot* ($p<0.05$) were considered significant, but

were accepted only upon manual inspection. Borderline hits were additionally verified by nanoES MS/MS.

1.2.7.4 Knowledge databases

Protein-protein interaction data were retrieved from the following knowledge databases: BioKnowledge database (includes YPD, PombePD, HumanPD, WormPD) available at Incyte Ltd. http://incyte.com (access provided by MPI CBG); CellZome Yeast Database available at http://yeast.cellzome.com (free access).

3. Results and Discussion.

3.1. Mass spectrometric analysis of gel separated proteins for proteomic applications.

Protein mixtures obtained in affinity purification experiments were separated by one-dimensional polyacrylamide gel electrophoresis prior to their identification by mass spectrometry. According to the established protocol [23] bands were visualized by staining with Coomassie Brilliant Blue or silver, excised from the gel slab, proteins were *in-gel* digested and unseparated mixtures of tryptic peptides subsequently analyzed by mass spectrometry. Efficiency of the *in-gel* digestion and comprehensive detection of peptides were the most important factors, which influenced the sensitivity and confidence of protein identification. This chapter is focused on how staining, storage and archiving of polyacrylamide gels affected tryptic digestion of protein bands and mass spectrometric detection of recovered peptides.

The following specific questions have been addressed here:

- How gel staining affects the yield of *in-gel* digestion products? How to adapt the *in-gel* digestion protocol to various methods of protein visualization?
- Does archiving and storage of gels affect *in-gel* digestion of proteins and subsequent mass spectrometry detection of peptides?
- How does long-time storage of archived gels affect peptide mass fingerprints of proteins embedded into polyarylamide matrix?
- What analytical strategy enables the most comprehensive characterization of *in-gel* protein digests by mass spectrometry?

3.1.1 A method for relative quantification of gel separated proteins

To provide consistent evaluation of how the sample preparation protocol affects the efficiency of *in-gel* digestion, it is important to accurately quantify the yield of individual digestion products, rather than to rely upon the qualitative assessment of peptide mass fingerprints or signal-to-noise ratios determined for particular peaks. Therefore we first developed a mass spectrometry-based method for the quantitative evaluation of the efficiency of *in-gel* cleavage of proteins and applied it for the comparison of the peptide recovery from gels stained by several commonly used protocols.

Peptides, which either incorporate amino acids enriched with stable isotopes, or are labeled by stable isotopes *via* a chemical or enzymatic reaction, are commonly employed as internal standards for the mass spectrometric quantification. Physicochemical properties of isotopically labeled peptides and unlabeled peptides from the analyte are almost identical, and this is important since both the amino acid composition and amino acid sequence strongly affect the intensity of detected peptide signals [125, 126]. However, because of the difference in their intact masses and masses of their fragment ions, they can be readily distinguished by MS and MS/MS methods.

3.1.1.1 Enzymatic labeling of tryptic peptides by ^{18}O-atoms

We employed enzymatic labeling of peptides with ^{18}O-atoms to produce internal standards for the quantification of proteins. First, an aliquot of the stock solution of a quantified protein with the known concentration was digested with trypsin in a buffer containing $H_2{}^{18}O$. Trypsin is a protease of choice for many proteomic applications that rely on the identification of proteins by mass spectrometry. First, the enzyme is very specific. It cleaves exclusively at the C-terminal site of lysine and arginine residues, if these residues are not followed by proline. Tryptic digestion yields predictable and highly reproducible sets of peptides, along with a relatively small number of autolysis products, whose masses can be subtracted from the mass spectrum. Second, masses of tryptic fragments are typically within a range of 500 – 2500 Da and can be determined by modern mass spectrometric instruments with better than 50 ppm mass accuracy [26].

Upon digesting of a protein in the buffer, which contains $H_2{}^{18}O$, tryptic peptides incorporate one or two ^{18}O-atoms into their C-terminal carboxyl groups [75, 127, 128] (Figure 7). Masses of labeled peptides are shifted by 2 and 4 Da respectively, comparing to the masses of native peptides. Comparison of peptide mass maps acquired from the digests of various standard proteins revealed that peptides with C-terminal arginine residues mostly incorporated two ^{18}O atoms ($2x^{18}O$ - peptides), whereas peptides with C-terminal lysine residues incorporated only one ^{18}O atom ($1x^{18}O$ - peptides) (Figure 8). Typically, less than 10% of the total amount of an arginine-containing peptide incorporates only one ^{18}O-atom [115].

Stability of $2x^{18}O$ - peptides was tested in a separate experiment by their incubation in 10 % formic acid in $H_2{}^{16}O$ for a long period of time. A stock solution of BSA was digested overnight with trypsin in $H_2{}^{18}O$ buffer and MALDI peptide mass fingerprint was acquired. An aliquot of the digest was dried down, re-dissolved in 10 % formic acid in $H_2{}^{16}O$ and then the solution was incubated at room temperature for several days. Peptide mass fingerprints of the mixture before and after the incubation were compared. We observed that $2x^{18}O$ - peptides underwent back-exchange of ^{18}O-atoms rendering $1x^{18}O$ –peptides and unlabeled peptides. However, for all examined peptides back-exchange was very slow and required several days before any noticeable alteration of the isotopic profile was detected. Routine probe preparation for MALDI peptide mapping that utilized 0.5% TFA or 5% formic acid as solvents did not decrease the content of ^{18}O-labeled standards and did not affect the accuracy of quantification.

Thus we concluded that enzymatic labeling of tryptic peptides with ^{18}O-atoms provided a simple and efficient method for the preparation of stable isotopically labeled peptide standards for mass spectrometric quantification experiments [129].

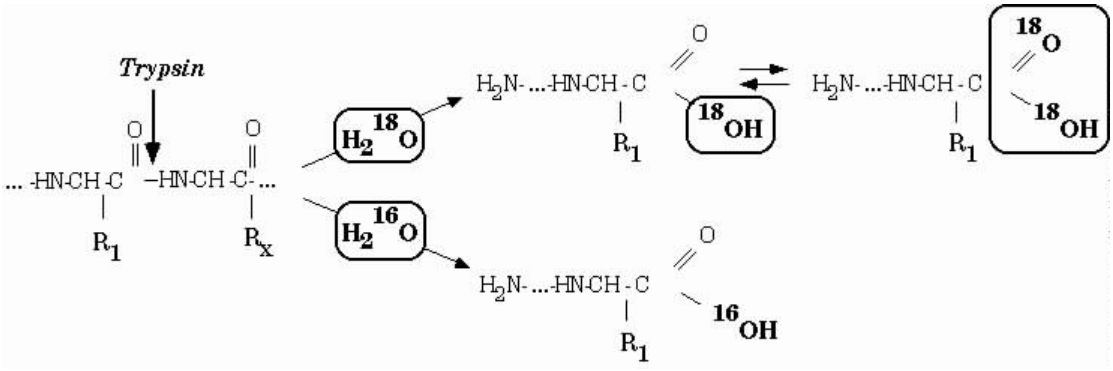


Figure 7 Enzymatic ^{18}O labeling of peptide standards for quantification. Upon digestion of a standard protein with trypsin in $H_2{}^{18}O$, ^{18}O atom is incorporated into C-terminal carboxyl groups of arginine and lysine residues. Trypsin catalyzes further exchange of another ^{16}O atom to ^{18}O atom in peptides with C-terminal arginine residues, whereas this does not occur in peptides with C-terminal lysine residues, which retain one ^{16}O atom in their carboxyl group. Therefore masses of Arg-containing peptide standards are shifted by 4 Da compared to the masses of quantified native peptides, which are produced by the protein cleavage in ^{16}O water [128].

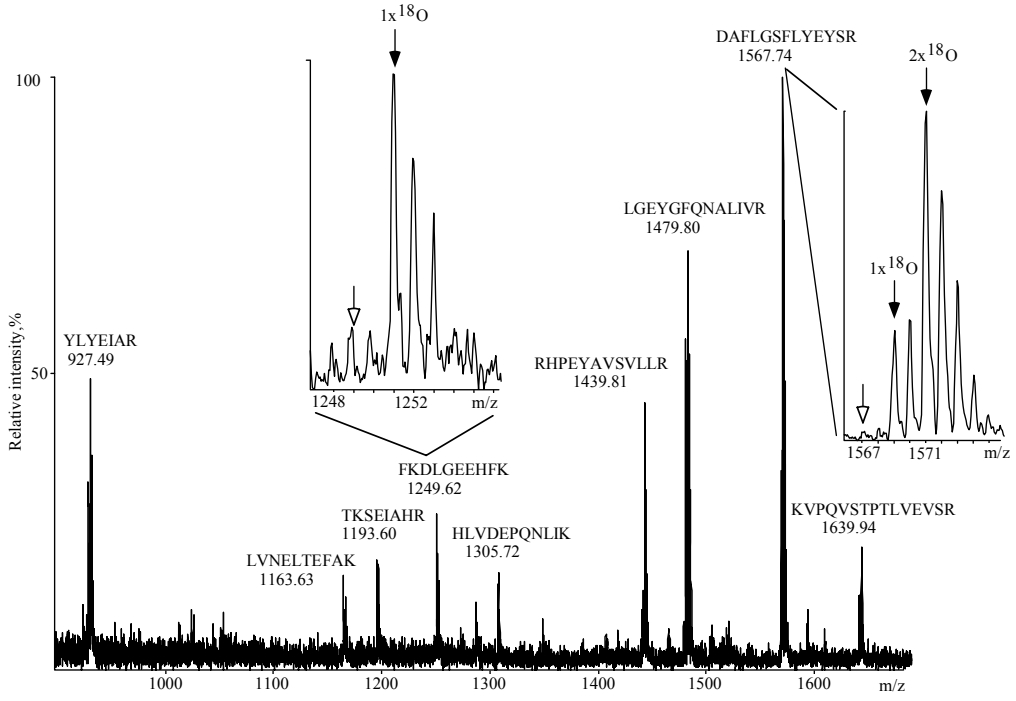


Figure 8 A part of the spectrum of the tryptic digest of BSA in a buffer containing $H_2{}^{18}O$. Peaks in the spectrum are designated with corresponding peptide sequences and *m/z* calculated for the unlabeled monoisotopic ions. Blow-outs demonstrate isotopic profiles typical for the peptide ions having arginine or lysine residues at their C-termini. The positions of the corresponding monoisotopic unlabeled ions are designated with unfilled arrows [129].

3.1.1.2 Calculation of the relative yield of *in-gel* digestion products

An aliquot of the mixture of ^{18}O – labeled peptides obtained by digesting a solution of the protein with known concentration (the internal standard) was mixed with an aliquot withdrawn from the experimental *in-gel* digest of the same protein that was performed in $H_2{}^{16}O$ (Figure 7) and the mixture was analyzed by MALDI TOF MS. MALDI peptide maps of the in-solution and *in-gel* digests of the same protein were similar, although the relative intensity of peptide peaks might be altered.

MALDI peptide mass fingerprints of tryptic digests are usually dominated by peptides containing arginine residues at their C-termini [125]. When produced by digestion of a protein standard in ^{18}O-water, these peptides appeared to be $2x^{18}O$ – labeled, and their molecular masses were increased by 4 Da (section 3.1.1.1). We used these peptides to estimate the yield of *in-gel* digestion. The relative concentration of a digestion product was calculated as a ratio of the intensity of the monoisotopic peak of the unlabeled peptide and the intensity of the monoisotopic peak of the corresponding $2x^{18}O$ -labeled peptide standard [129, 130] (Figure 9). MALDI measurement of each aliquot from each sample was repeated 4-5 times and quantification results were averaged.

Profiles of the isotopic clusters could be more accurately deconvoluted using isotopic ratios calculated from the peptide elemental composition [115, 130, 131]. Deconvolution is important to determine the yield of peptides with C-terminal lysines, which mostly produced $1x^{18}O$-labeled forms upon digestion of proteins in $H_2{}^{18}O$. However, in peptide mass finferprints these peptides are usually less abundant than peptides with C-terminal arginine residues. Since chemical noise strongly affects the accuracy of deconvolution, we found that accounting for peptides with C-terminal lysine residues only decreased the accuracy of measurements and therefore was not employed in this study.

Linearity of the calibration curve was tested by analyzing series of samples obtained by successive dilution of an aliquot withdrawn from the *in-gel* digest of 1 pmol of BSA. Relative concentrations calculated for various peptides were found linear at least over 1:5 dilution range and were affected by chemical noise at larger dilution ratios.

We observed that the accuracy of quantification might be compromised if peaks of the analyte or internal standards happen to overlap with another peptide

peak, originated from a measured protein or from a common protein contaminant, such as human or sheep keratins or antibodies [132] (Figure 10).

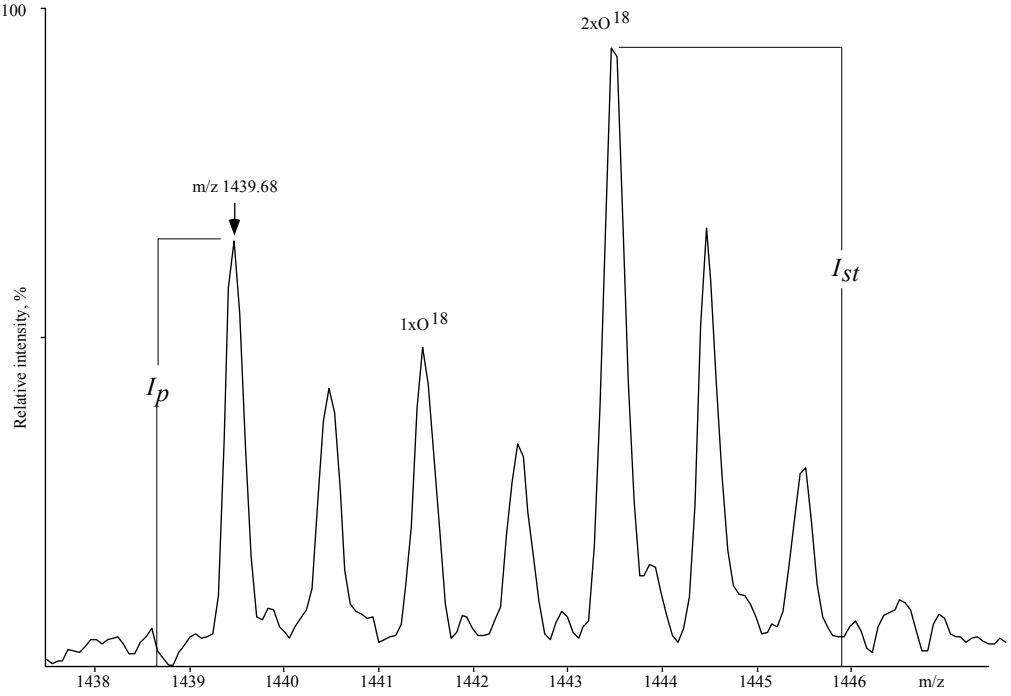
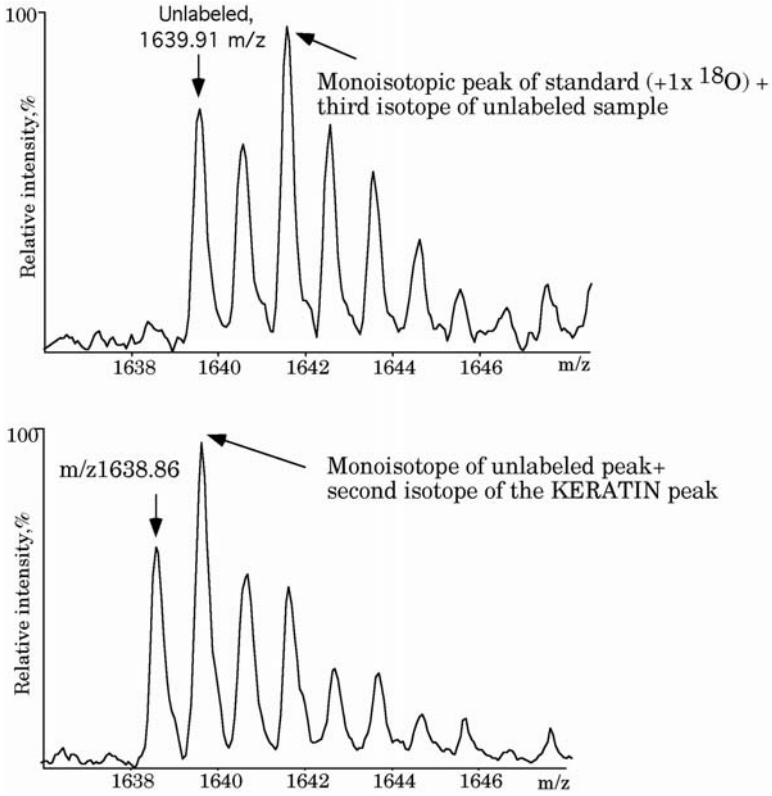


Figure 9 Calculation of the relative concentration of peptides. A blow-out of the isotopic cluster of the peptide peak with *m/z* 1439.93 from the tryptic digest of BSA (RHPEYAVSVLLR). The monoisotopic peak of the unlabeled peptide is designated with a filled arrow. The peak of the isotopically labeled peptide that incorporated two ^{18}O atoms ($2x^{18}O$) was used as an internal standard. It was estimated by comparing the computed and detected isotopic profiles that the combined contribution of the intensity of the fourth isotopic peak of the unlabeled peptide, and of the second isotopic peak of $1x^{18}O$-peptide into the intensity of the monoisotopic peak of the $2x^{18}O$-peptide did not exceed 10 % and was at the level of background noise. The relative concentration of the unlabeled peptide (R_c) was calculated as $R_c = I_p/I_{st}$, where I_p stands for the intensity of the peptide peak and I_{st} for the intensity of the peak of the standard [129].



Figures 10 By overlapping with the quantified peptide, a peptide from the protein contaminant (keratin) can affect the results of relative quantification. Upper panel: isotopic profile of the mixture of ^{18}O-labeled and native BSA tryptic peptide with *m/z* 1639.91. Lower panel: isotopic profile has changed when the sample was contaminated by keratin. Monoisotopic peak of the abundant keratin peptide was detected at *m/z* 1638.86 [128].

3.1.1.3 How staining of polyacrylamide gels affects the yield of tryptic peptides?

Using the method described in section 3.1.1 the effect of gel staining on the yield of tryptic peptides was examined by analyzing *in-gel* digests of bands containing 1 pmol of BSA, which were visualized by Coomassie Brilliant Blue, silver [23] and zinc-imidazole (negative) [114] staining. We further tested if the recovery of peptides from silver stained gels could be improved by a pre-digestion destaining of protein bands as was suggested previously by Gharahdaghi *et al* [113].

Several protein bands were in parallel digested with trypsin and the digests were analyzed by MALDI TOF. A similar profile of tryptic peptides was detected in

each of these samples independently from the method of gel staining (Figure 11) and relative concentrations were determined for the five most intense peptides with C-terminal arginine residues (Table 1).

Table 1

Sequences and *m/z* of BSA tryptic peptides employed in the relative quantification of the yield of *in-gel* digestion.

Peak number	*m/z*	Sequence
1	927.49	YLYEIAR
2	1439.81	RHPEYAVSVLLR
3	1479.80	LGEYGFQNALIVR
4	1567.74	DAFLGSFLYEYSR
5	1639.94	KVPQVSTPTLVEVSR

Small variation of the relative concentration of individual peptides was observed in *in-gel* digests of bands stained with different methods. However, none of the staining methods provided significantly better recovery of peptides compared to other methods (Table 2).

Also the number of detected peptide peaks and the yield of corresponding peptides did not increase substantially, if silver stained bands were destained prior to the digestion (Table 3). We note, however, that the relative concentration of all peptides in the digests of silver stained bands, which were treated with trypsin directly (*i.e.* washing steps as well as destaining, reduction and alkylation were omitted), was dramatically lower (Table 3).

We observed that reduction and alkylation did not influence the recovery of peptides, which do not contain cysteine residues and therefore, to speed up protein identification, they could be omitted [115, 133]. However, cysteine-containing peptides were completely absent in the digest of a non-reduced protein (Figure 12). Since confidence of the protein identification by MALDI peptide mapping depends on the number of peptides matched to the protein sequence, loss of several peptides hampered or even made impossible the identification of proteins that rendered only a few tryptic fragments (for example, low molecular weight proteins, membrane proteins, etc.).

Independently from the gel staining protocol and pre-digestion treatment of protein bands, the number of detected peptides was always sufficient for the

unambiguous identification of BSA upon searching a database. We thus concluded that abundant protein bands can be identified upon tryptic digestion directly after their visualization and washing with the digestion buffer. However, low abundant silver stained bands should be reduced and alkylated prior to the digestions in order to increase the number and yield of detected peptides.

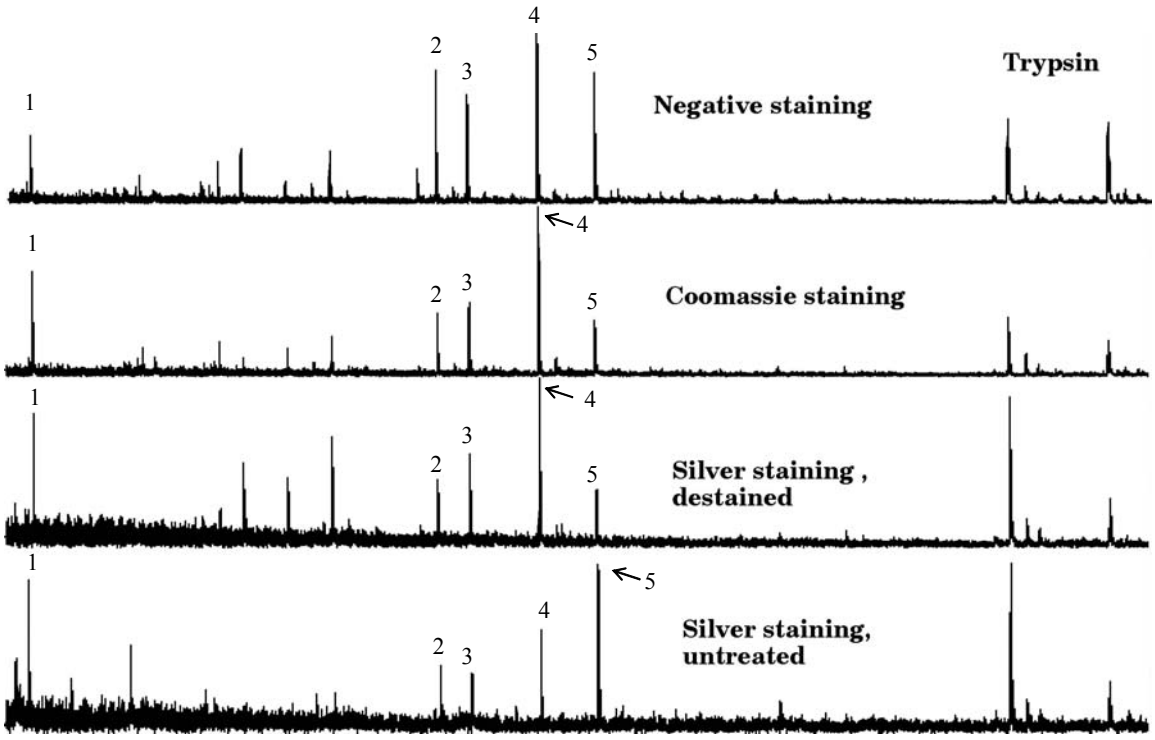


Figure 11 MALDI peptide maps of *in-gel* tryptic digests of bands containing 2 pmol of BSA, which were visualized according to various staining protocols. From the top: Zn-imidazole (negative) staining; Coomassie Brilliant Blue staining; silver staining with destaining and without destaining. Peaks are numbered as in Table 1 [128].

Table 2

Relative concentration of tryptic peptides of BSA in *in-gel* digests of bands stained by various methods

Staining method	Relative concentration* (%)				
	m/z 927.49	*m/z* 1439.81	*m/z* 1479.80	*m/z* 1567.74	*m/z* 1639.94
Silver, with reduction and alkylation	100	100	100	100	100
Coomassie	105	136	74	95	121
Zn/Imidazole	117	61	74	79	100

* Relative concentrations of peptides were normalized to their concentrations in the digests of silver-stained bands.

Table 3.

Relative concentration of peptides recovery from silver-stained gels

Sample preparation method	Relative concentration[a] (%)				
	m/z 927.49	*m/z* 1439.81	*m/z* 1479.80	*m/z* 1567.74	*m/z* 1639.94
With destaining, reduction and alkylation	100	100	100	100	100
With destaining only	95	110	93	97	98
With reduction and alkylation	115	104	102	83	94
Untreated[b]	<5	12	14	6	19

[a] Relative concentrations of peptides were normalized to the concentrations in digests of destained, reduced and alkylated bands.

[b] Pre-digestion washing, destaining, reduction and alkylation were omitted

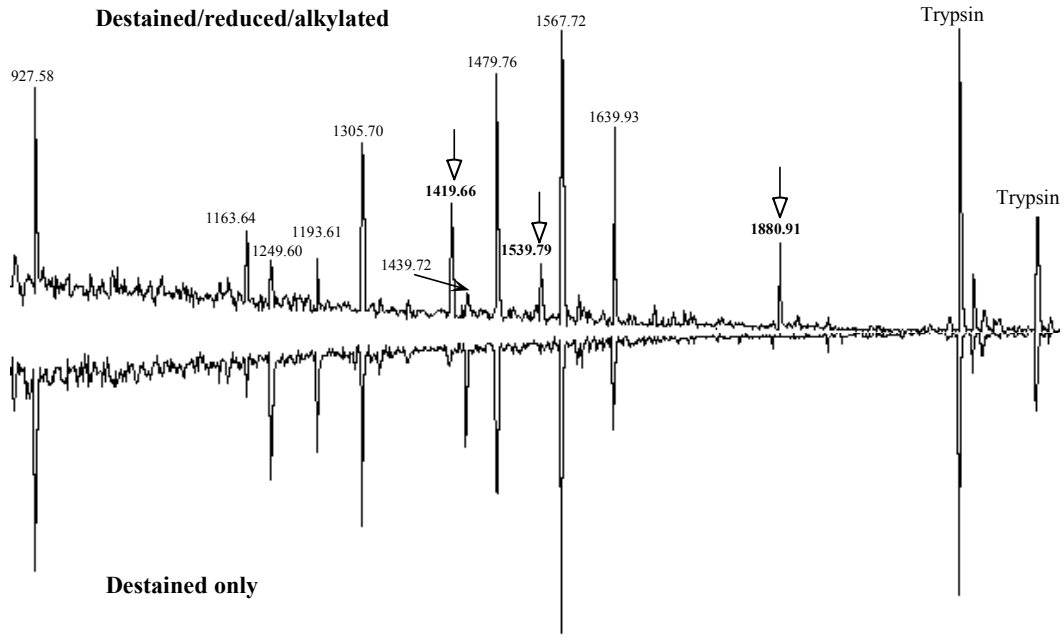


Figure 12. Comparison of the peptide maps of the silver stained bands processed using destaining, reduction and alkylation (the upper spectrum) and using only destaining (the lower spectrum). Peaks are designated with corresponding *m/z*, peptide sequences are presented in Figure 2. Three intense peptide peaks (designated with unfilled arrows) matching cysteine-containing peptides from BSA were additionally detected after reduction and alkylation. Corresponding peptide sequences are: *m/z* 1419.66 SLHTLFGDELCK; *m/z* 1539.79 LCVLHEKTPVSEK; *m/z* 1880.91 RPCFSALTPDETYVPK. C stands for cysteine-S-acetamide residues [129].

3.1.2 Comprehensive characterization of complex *in-gel* digests by mass spectrometry

MALDI MS is one of the most frequently used technologies in proteomics because of the low femtomole sensitivity, high throughput and simple and robust sample preparation. However, confident identification of minor components in mixtures, identification of low molecular weight proteins, as well as of hydrophobic or very hydrophilic proteins often could not be achieved by MALDI MS alone. This section demonstrates how the analysis by complementary mass spectrometric techniques combined with smart database searches expands the scope of MALDI-based proteomics.

3.1.2.1 Characterization of proteins by a combination of complementary mass spectrometric techniques

To demonstrate how complementary mass spectrometric methods enable in-depth characterization of protein digests, we analyzed 40kDa Coomassie stained spot excised from a two-dimensional polyacrylamide gel as a case study. This spot contained a human protein co-immunoprecipitated by polyclonal antibodies raised against CD45 receptor (isolation was peformed in the laboratory of Dr. B. Schraven in DKFZ, Heidelberg [140]). The spot was *in-gel* digested and the recovered mixture of tryptic peptides was analyzed by three mass spectrometric techniques – MALDI MS, MALDI QqTOF MS and nanoES MS/MS. The comparison of peptide patterns detected by these three methods is provided in Table 4.

MALDI peptide mapping (Figure 13A) identified the sample as a mixture of two human proteins, both having a molecular weight close to 40 kDa: P04901 transducin beta-chain 1 (gene name *GNB1*) and P11016 transducin beta-chain 2 (gene name *GNB2*), which produced partly overlapping peptide mass maps. Altogether, 16 peptide masses were matched to the sequences of transducins. Almost all intense peptide peaks detected in the MALDI spectrum could be assigned to either of these two proteins, with exception of two intense peptide peaks having *m/z* 1728.78 and 1770.81 (labeled as T17 and T18 in Figure 13A).

To determine the identity of these two prominent peaks nanoES MS/MS analysis was applied. NanoES MS/MS confirmed the presence of GNB1 and GNB2 (Figure 13B, Table 4) and suggested that at least some of the unmatched peaks can be explained by protein polymorphism. Sequences of two peptide precursors T7 and T8 (Table 4) differed from the sequence of the corresponding peptide in a database (T6) by a single amino acid residue. However, the observed profile of peptide peaks looked strikingly different from the MALDI map, and no apparent candidates for T17 and T18 were found.

MALDI QqTOF MS typically requires a “cold” matrix like 2,5-dihydroxybenzoic acid (DHB) for best results [35, 36]. Thus, it is not surprising that relative intensities of peptide ions in the spectra obtained by conventional MALDI and by MALDI QqTOF MS were somewhat different, although in general similar profiles of peptide ions were observed (Figure 13A and C). Importantly, the unidentified peptide ions (*m/z* 1728.854 and *m/z* 1770.874) were also detected as

intense peaks. Upon partial manual interpretation and error-tolerant database searching, tandem mass spectra acquired from T17 and T18 (Figure 13D) revealed that these peptides originated *via* *N*-terminal processing of GNB1 and GNB2, respectively. The *N*-terminal methionine residues had been removed and the resulting peptides were acetylated.

Thus despite the complexity of a mixture of two highly homologous *N*-terminally processed proteins, almost all prominent peptide ions in the peptide map were assigned to the protein sequences.

This example illustrates the need for a combination of mass spectrometric techniques to fully elucidate the composition of protein samples. NanoES MS and MALDI MS detect complementary peptide patterns, and their combination increases confidence in protein identification [134, 135]. However, applying two ionization methods to the analysis of the same protein digest limits the throughput, which is an important factor for medium and large scale proteomic projects.

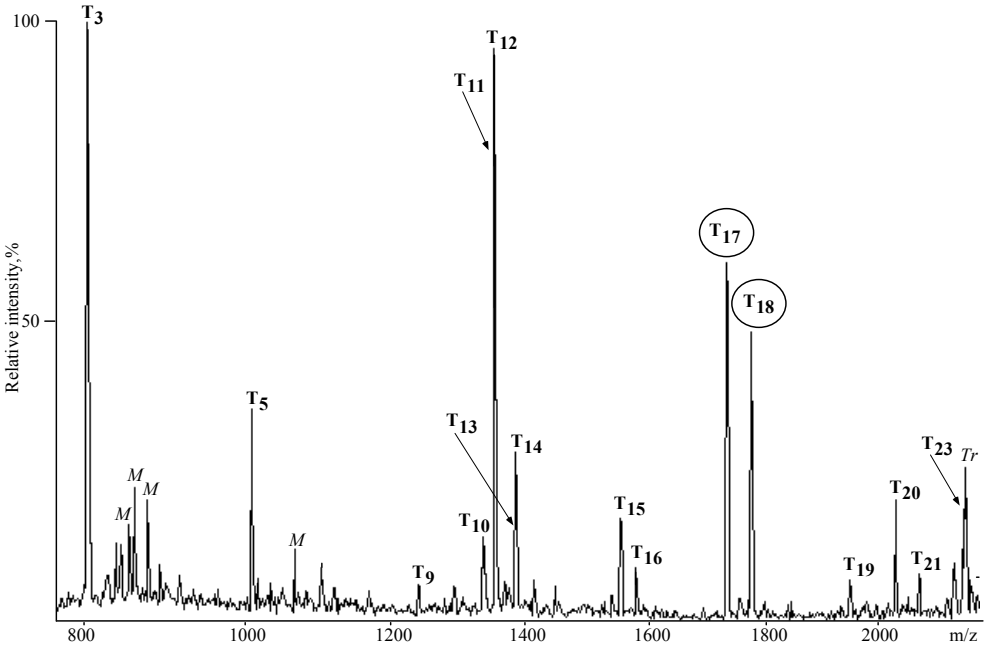
Table 4

Identification of peptides in the *in-gel* digest of 40 kDa human protein by complementary mass spectrometric techniques – MALDI MS, nanoES MS/MS and MALDI QqTOF MS

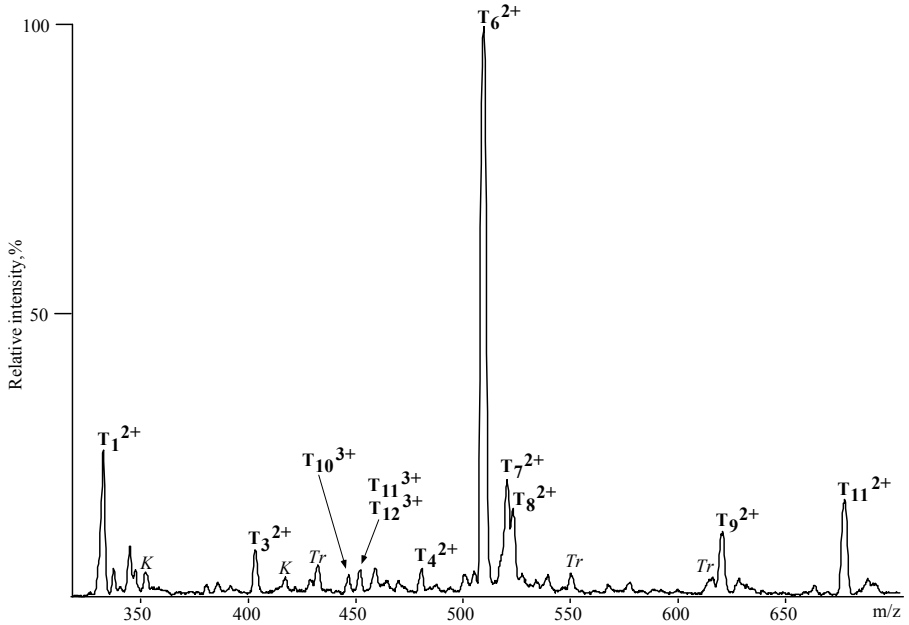
Peptide	Residues	Sequence	MALDI	MALDI QqTOF	ES MS	GNB 1 or 2
T_1	252 - 256	LFDLR		•	•	GNB1/2
T_2	210 - 214	LWDVR		•		GNB1/2
T_3	90 - 96	VHAIPLR	•	•	•	GNB1/2
T_4	179 - 197	(M)SLSLAPDTR			•	GNB1
T_5	305 - 314	AGVLAGHDNR	•	•		GNB1/2
T_6	69-78	LLVSASQDGK			•	GNB1/2
T_7	69 - 78	LLVSASQ**H**GK			•	—
T_8	69 - 78	LLV**D**ASQDGK			•	—
T_9	198 - 209	LFVSGAC(acr)DASAK	•		•	GNB1
T_{10}	58 - 68	IYAMHWGTDSR	•	•	•	GNB1/2
T_{11}	58 - 68	IYAM(ox)HWGTDSR	•	•	•	GNB1/2
T_{12}	79 - 89	LIIWDSYTTNK	•	•	•	GNB1/2
T_{13}	58 - 68	IYAM(ox)HW(ox^2)GTDSR	•			GNB1/2
T_{14}	79 - 89	LIIW(ox^2)DSYTTNK	•			GNB1/2
T_{15}	138 - 150	ELAGHTGYLSC(acr)C(acr)R	•	•		GNB1
T_{16}	138 - 150	ELPGHTGYLSC(acr)C(acr)R	•	•		GNB2
T_{17}	2 - 15	(N-acetyl)SELDQLRQEAEQLK	•	•		GNB1
T_{18}	2 - 15	(N-acetyl)SELEQLRQEAEQLK	•	•		GNB2
T_{19}	24 - 42	AC(acr)GDSTLTQITAGLDPVGR	•			GNB2
T_{20}	24 - 42	AC(acr)ADATLSQITNNIDPVGR	•	•		GNB1
T_{21}	23 - 42	KAC(acr)GDSTLTQITAGLDPVGR	•			GNB2
T_{22}	79 - 96	LIIWDSYTTNKVHAIPLR		•		GNB1/2
T_{23}	284 - 301	KAC(acr)ADATLSQITNNIDPVGR	•	•		GNB2

Detected peptides are designated with filled circles. C(acr) stands for cysteine-S-acrylamide; M(ox) stands for methionine sulfoxide; W(ox^2) stands for doubly oxidized tryptophane.

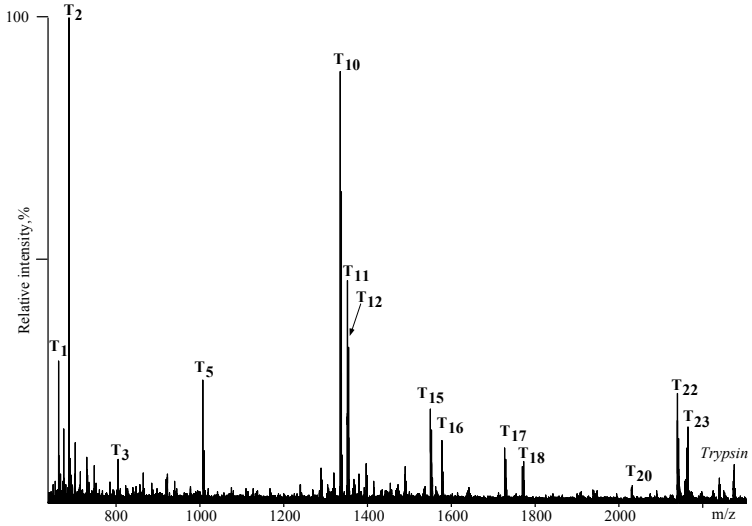
A


100
50
Relative intensity,%
T3
T12
T11
T17
T18
T5
T13
T14
T23
Tr
M
M
M
T20
T15
T10
M
T16
T9
T19
T21
800
1000
1200
1400
1600
1800
2000
m/z

B


100
50
Relative intensity,%
T6 2+
T1 2+
T10 3+
T11 3+
T12 3+
T7 2+
T8 2+
T11 2+
T3 2+
T9 2+
K
K
Tr
Tr
Tr
T4 2+
350
400
450
500
550
600
650
m/z

C



D

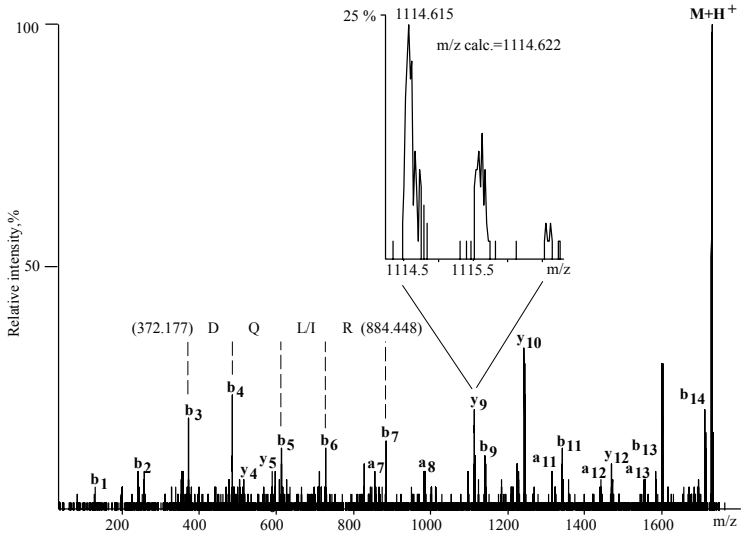


Figure13. (A) MALDI peptide map acquired from the digest of p40. Peptide ions labeled in the panel were matched to the sequences of GNB1 and/or GNB2 (Table 6). Peaks designated with T17 and T18 (circled) were not matched. (B) NanoES spectrum of the same digest. All peaks labeled in the spectrum were assigned. (C) MALDI-QqTOF spectrum of the digest. Peaks of the peptides T17 and T18 were observed in the spectrum and their MS/MS spectra were subsequently acquired. (D) The tandem mass spectrum acquired from the precursor ion with *m/z* 1728.853. A stretch of the peptide sequence was deduced from the series of b-ions and assembled into a sequence tag together with masses of the corresponding fragments. Upon error tolerant searching the identity of T17 was established [134].

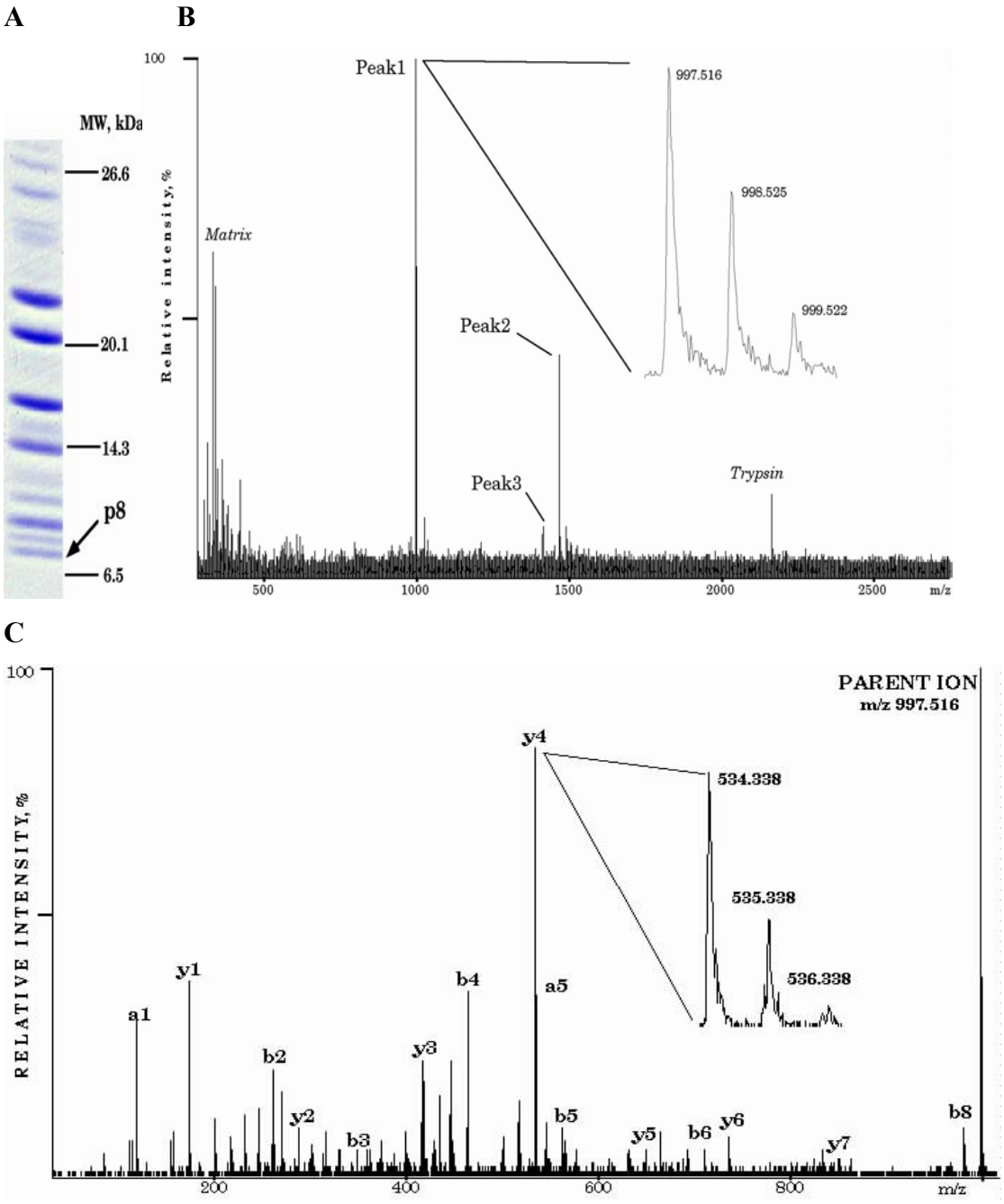
3.1.2.2 Mass spectrometric identification of low molecular weight proteins

Proteins with molecular weights lower than 20 kDa often produce only a few peptides upon their cleavage with trypsin. Typically, only one to four peptides are detected in MALDI spectra of their digests. Since MALDI peptide mass fingerprinting heavily relies on the number of detected and matched peptide masses, statistically confident identification of these proteins might be problematic even if they were purified in the amount of several picomoles. The analysis is becoming even more complex if the sample is a mixture of several small proteins, and each protein component is represented only by a single peptide. This problem is commonly encountered in the analysis of almost any band excised from a one-dimensional gel in the region below ca 20 kDa. Along with common (although poorly reproducible) background of keratin peptides and (sometimes) antibodies, these bands often contain truncated fragments of abundant proteins with higher molecular weight. Therefore increasing the specificity of database searches by imposing additional constraints on protein modifications, species of origin, molecular weight and mass accuracy could not address the complexity of peptide mass patterns.

This section demonstrates that a combination of MALDI MS analysis with complementary mass spectrometric technique (MALDI Qq TOF or ESI MS/MS) provides an alternative approach for identification of gel-separated low molecular weight proteins.

The analysis of an intense Coomassie-stained band of *S. cerevisiae* protein with apparent molecular weight of approximately 8 kDa (purified by the group of Dr. B. Seraphin (EMBL, Heidelberg)) is presented as a case study (Figure 14A). When the *in-gel* tryptic digest of this band was analyzed by MALDI peptide mass fingerprinting, only 3 peptide ions were detected and database search did not result in any plausible identification. Therefore, tryptic peptides were extracted and analyzed on a prototype MALDI QqTOF instrument. As anticipated, the same three peptide masses were detected and, despite of better than 20 ppm mass accuracy of the instrument, database search was also unsuccessful (Figure 13B). Tandem mass spectrum was acquired for the most intense precursor peak with *m/z* 997.514 (Figure 14C). Masses of 34 fragment ions and the mass of the precursor were searched against a comprehensive protein database by MS-Tag program at the web server of Mass

Spectrometry Facility at UCSF and no restrictions on species of origin and protein molecular weight were applied. The search produced a single confident hit – the peptide FNSDVFLR originating from the budding yeast protein Lsm6p (YDR 378c) having the molecular weight of 13.7 kDa. The two remaining peptide masses were retrospectively matched to peptides from its sequence. One of the two peptides was the very C-terminal non-tryptic fragment containing oxidized methionine (Figure 14D). All three peptides flanked the C-terminal part of the protein that covered approximately a quarter of its full-length sequence, although the remaining ¾ additionally comprised 13 potential tryptic cleavage sites. Together with the observation that the apparent molecular weight was lower than anticipated (8 kDa instead of 13 kDa) this indicated that its sequence was truncated from the N-terminus.



D. Chromosome IV Cosmid 948; Gene name YDR378C, MW 13.6kDa

MPNKQRRSNP FNNAQDAILT RKYGIAQYKQ NKQRLLVMSG KASTEGSVTT EFLSDIIGKT VNVKLASGLL
YSGRLESIDG FMNVALSSAT EHYESNNNKL LNKFNSDVFL RGTQVMYISE QKI

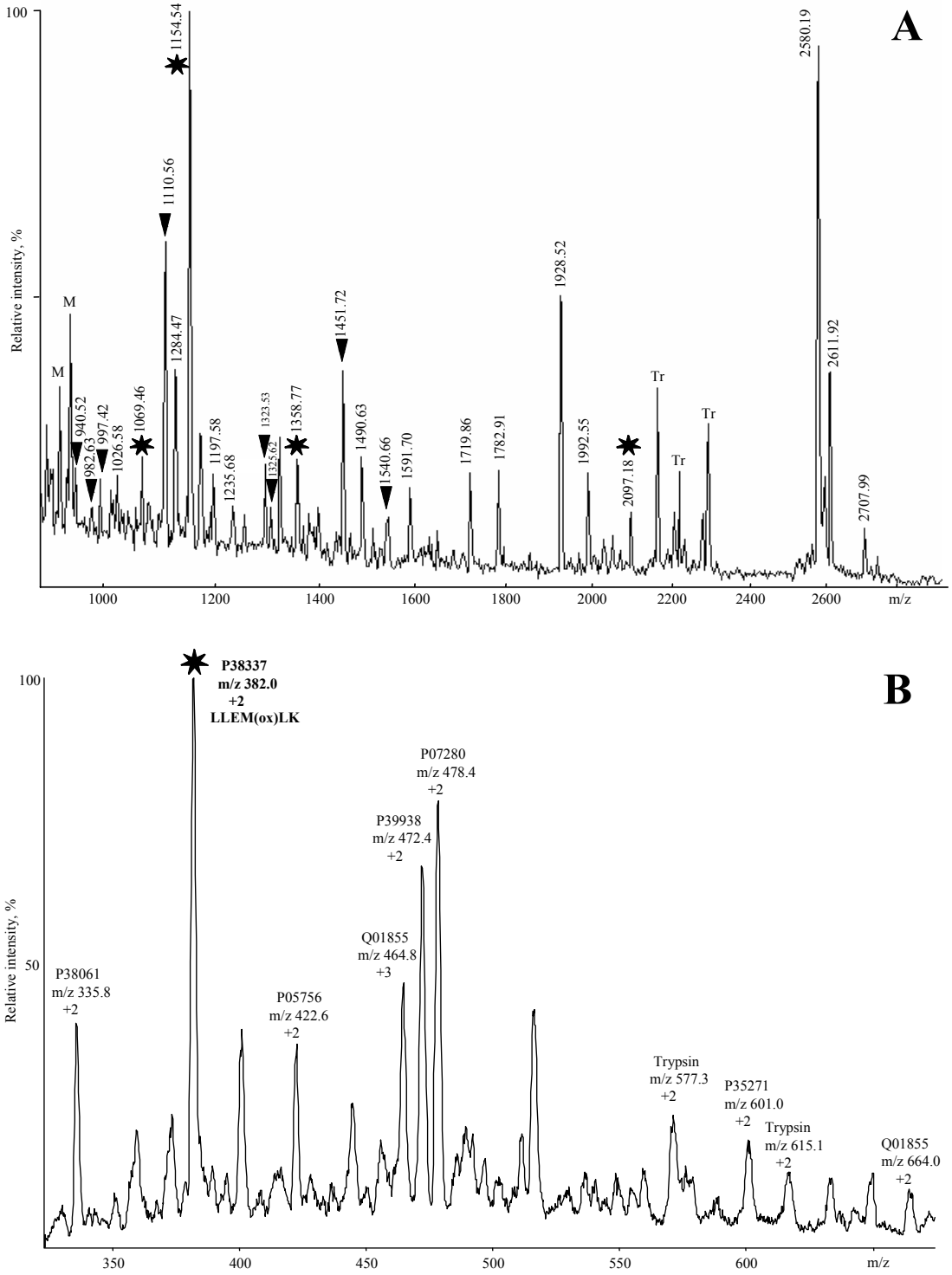
	Sequence	Position
Peak1	FNSDVFLR	104-111
Peak2	LLNKFNSDVFLR	101-111
Peak3	GTQVM(ox)YISEQKI	112-123

Figure 14. Identification of a low molecular weight yeast protein by MALDI QqTOF MS. (A) Coomassie-stained band with a molecular weight of about 8 kDa (p8) was excised from a 1D gel. (B) Peptide mass fingerprint of *in-gel* tryptic digest of p8; (C) MS/MS spectrum acquired form the precursor ion with *m/z* 997.516; (D) All peptides detected in the mass fingerprint (panel A) match the C-terminal part of the sequence of Lsm6p.

Alternatively, digests of low molecular weight proteins could be analyzed by a combination of mass spectrometric techniques, which utilize different types of ionization and could therefore increase the coverage of corresponding protein sequences, as demonstrated by the analysis of the Coomassie-stained band of the 15 kDa budding yeast protein purified in the laboratory of Dr. A.F. Stewart (EMBL, Heidelberg). The band was excised from a one-dimensional gel, *in-gel* digested and the peptide mixture analyzed by MALDI MS [136]. Although the apparent molecular weight of the band was low, 25 prominent peptide peaks were detected in the spectrum (Figure 15A). Database search with their masses resulted in a single confident hit – the 15 kDa 40S ribosomal protein S24 (accession number P26782); MOWSE score of the hit was 115, whereas the suggested threshold of statistical confidence was 52). Eight peptide ions were matched with better than 100 ppm mass accuracy and covered more than 50% of the protein sequence. However, most intense peptide peaks in the spectrum remained unaccounted for, and further database searches with masses of unmatched ions did not result in any more hits, indicating that other yet unidentified protein(s) might be present in the sample. The digest was then analyzed by nanoES MS/MS, and another six ribosomal proteins, each of which matched by a single unique sequenced peptide, were identified (Figure 15B). Some of these proteins were previously found among non-confident hits of MALDI peptide mass fingerprinting; however, surprisingly, none of the sequenced peptides identified S24 protein, the top hit. A single peptide sequence deduced from the spectrum acquired from a doubly charged ion with m/z 382.0 matched the 15 kDa yeast protein YBR258c (Figure 15C). However, this hit could not be considered as confident since the retrieved sequence was short and degenerate. It is also known that large multiply charged peptides often undergo partial orifice fragmentation yielding abundant singly and doubly charged *y*-ions and therefore database searches with such sequences should not rely upon the cleavage specificity of trypsin. In fact, the peptide sequence (Leu/Ile)(Leu/ Ile)Glu(Met(ox)/Phe)(Leu/Ile)(Lys/Gln) hit more than 200 proteins in a comprehensive database, including six proteins from the budding yeast. Retrospectively the MALDI MS map revealed that masses of another four peptides matched the sequence of YBR258c, however none of them matched other yeast protein candidates. It was therefore concluded that although neither MALDI MS nor nanoES MS/MS vouched for the unambiguous identification, a combination of the

two techniques produced a confident hit. Later biological experiments independently confirmed the presence of the YBR258c protein in the sample [137].

The case studies presented above demonstrated that the identification of small proteins is difficult and often ambiguous, even if complete and accurate protein sequences are present in a protein database. Careful inspection of combined data obtained by complementary mass spectrometric techniques increased the completeness of the sample characterization and confidence of protein identification.


A
100
Relative intensity, %
1154.54
1110.56
2580.19
M
M
1284.47
1451.72
1928.52
2611.92
940.52
997.42
1069.46
1026.58
1323.53
1358.77
1197.58
1235.68
1490.63
1540.66
1591.70
1719.86
1782.91
1992.55
2097.18
2707.99
Tr
Tr
Tr
1000
1200
1400
1600
1800
2000
2200
2400
2600
m/z
B
100
50
Relative intensity, %
P38337
m/z 382.0
+2
LLEM(ox)LK
P07280
m/z 478.4
+2
P39938
m/z 472.4
+2
Q01855
m/z 464.8
+3
P38061
m/z 335.8
+2
P05756
m/z 422.6
+2
Trypsin
m/z 577.3
+2
P35271
m/z 601.0
+2
Trypsin
m/z 615.1
+2
Q01855
m/z 664.0
+2
350
400
450
500
550
600
m/z

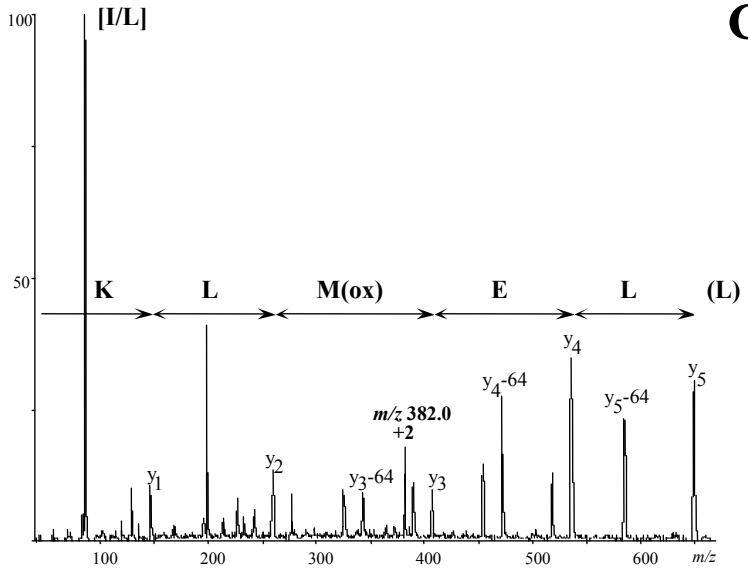


Figure 15. Mass spectrometric identification of a low molecular weight yeast protein. (A) MALDI MS peptide map of the tryptic digest of the 20kDa protein band. Peaks of autolysis products of trypsin are designated with Tr, matrix peaks with M. Peptide peaks, which were retrospectively matched to YBR258c, are designated with asterisks; peaks matched to the ribosomal protein S24 are designated with filled triangles. (B) NanoES mass spectrum of the same digest acquired in precursor ion scanning mode with the selected fragment ion at *m/z* 86. Peaks designated with *m/z* and charges were fragmented, and MS/MS spectra were matched to the sequences of tryptic peptides from various ribosomal proteins (see accession numbers at the corresponding peaks). The peak designated with the asterisk was the only peptide ion that matched yeast protein YBR258c. (C) tandem mass spectrum acquired from the peptide ion with *m/z* 382.0 and the candidate peptide sequence, deduced by considering mass differences between y-ions [136].

3.1.3 Mass spectrometric identification of proteins from archived polyacrylamide gels

Biochemical experiments often produce polyacrylamide gels containing potentially interesting proteins. However, if adequate protein identification capacities were missing, or if genomic sequencing of the organism was still in progress and only a limited number of sequences was available, such gels were typically dried on a plastic or paper support and kept for years in folders as reference images for visual inspection of changes in protein profiles. However, archived gels are a valuable resource for proteomics, since the identification of proteins from such gels might alleviate the need to repeat protein purification experiments. It has been demonstrated that proteins from archived gels stored for a long period of time at room temperature could, in principle, be identified by *in-gel* trypsinolysis followed by MALDI peptide mass mapping [138, 139]. However, no systematic studies on this subject were performed and three key questions remained open:

- Does archiving of gels affect the yield of *in-gel* digestion, the recovery of peptides from a gel matrix and, subsequently, the sensitivity of mass spectrometric identification of proteins?
- Could archived proteins be modified or damaged during storage of gels?
- How could these modifications be accounted for so that the confidence of protein identification would be preserved, and facile characterization of protein polymorphism would be possible?

Using the quantification method described in section 3.1 and a combination of complementary mass spectrometric techniques, *in-gel* digests of proteins from archived gels were subjected to rigorous qualitative and quantitative examination [134].

3.1.3.1 *In-gel* digestion of proteins from archived gels

Two archived gels obtained in independent projects were examined. The first gel contained human proteins immunoprecipitated by polyclonal antibodies against CD45 receptor [140]. Proteins were separated by two-dimensional gel electrophoresis and visualized by Coomassie staining. The gel had been subsequently dried between two sheets of a cellophane film and stored for more than eight years in a folder at

room temperature. 14 most prominent spots were selected for the analysis (Figure 16A).

Another one-dimensional gel, containing a chicken protein p80 co-purified with the bait by monoclonal antibodies against INCENP protein [141] was stained with Coomassie, dried on a paper support and stored at room temperature for more than 15 years before the analysis described here (Figure 16B).

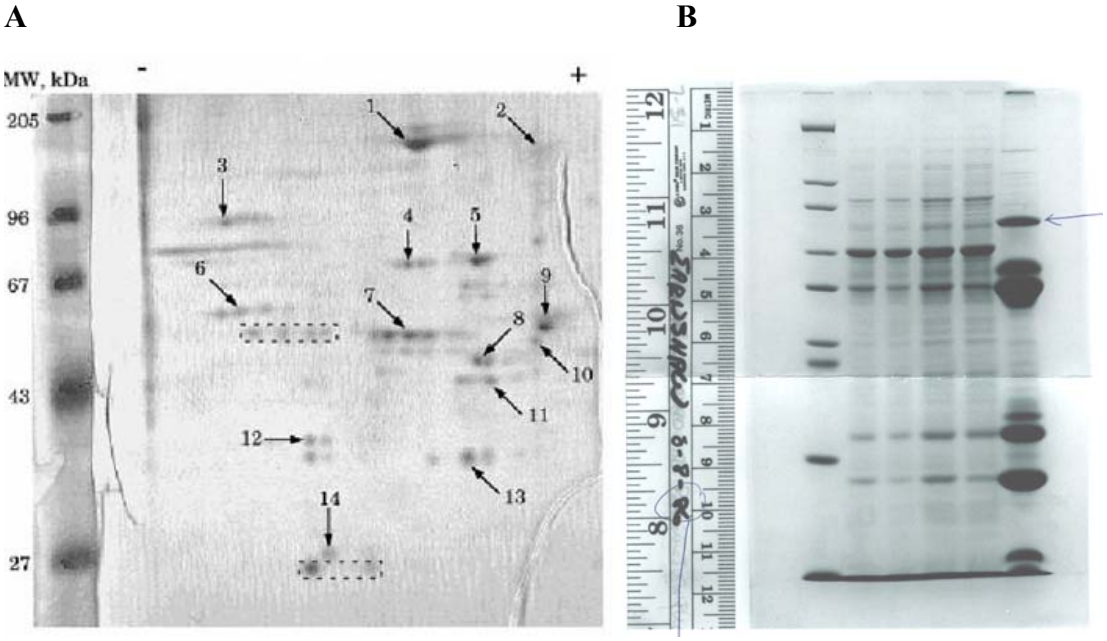


Figure 16 Images of the two Coomassie-stained archived gels analyzed in this work. (A) The gel was dried between cellophane sheets. Protein spots labeled with arrows were analyzed by mass spectrometry. Spots originating from the antibodies used in the immunoaffinity purification are boxed [134]. (B) The gel was dried on a paper support. The analyzed band of p80 protein is marked by arrow (kindly provided by Dr. W. Earnshaw, University of Edinburg).

No substantial changes in the conventional *in-gel* digestion protocol [23] were required to analyze proteins from archived gels. To avoid crushing of a gel slab during re-hydration, spots from both dried gels were excised first and then gel plugs were re-hydrated separately. After a brief incubation in water, pieces of cellophane could be easily detached from the gel plug. Importantly, pieces of dust, threads and traces of human fingerprints (frequent sources of keratin contamination) remain at the

surface of the film during re-hydration and can be subsequently removed together with the film after the re-hydration step is completed. For gels dried on a paper support, it is often difficult to remove paper debris from the reaction mixture, but residual paper filaments do not noticeably affect the recovery of digestion products. However, they complicate handling of samples because of frequent clogging of the pipette tips and they are a potential source of keratin contamination. Also one surface of a slab remains uncovered, if the gel was dried on a paper support and therefore the risk of contaminating protein samples during gel storage and handling increases.

3.1.3.2 Qualitative and quantitative evaluation of patterns of *in-gel* digestion products rendered from fresh and archived gels

To examine if archiving affects the yield of *in-gel* digestion products we further applied the quantification method described in section 3.1.1. We compared the yield of tryptic peptides obtained by *in-gel* digestion of Coomassie-stained bands of the model protein BSA excised from a freshly prepared "wet" gel and from a gel, which had been dried on cellophane support 48 h prior to digestion.

Two pmoles of BSA were loaded onto separate lanes of a one-dimensional polyacrylamide gel. After electrophoresis the gel was stained with Coomassie Brilliant Blue and then cut into two parts, each containing several BSA bands. One part was dried on a plastic support using a GelAir dryer and the other part was kept in water. Two BSA bands were excised both from the "wet" and archived parts of the gel and *in-gel* digested in parallel. ^{18}O-labeled tryptic peptides of BSA were employed as internal standards for quantifying the *in-gel* digestion peptide products. Relative yields of six tryptic peptides presented in Table 5 suggested that peptide recovery was not affected by gel archiving.

To investigate whether prolonged storage affected the pattern of digestion products, MALDI peptide mass maps were acquired from an *in-gel* digest of the BSA band excised from a freshly prepared gel and from a digest of spot 4 from the archived gel (Figure 16A), which was also identified as BSA (Table 7). Both spectra were acquired on the same instrument and samples were prepared using similar *in-gel* digestion and probe preparation recipes. In both spectra only a few minor peaks did not match BSA sequence. Almost identical pattern of peptides, which, however, did

not contain methionine and cysteine residues, was observed. At the same time many peptides containing *S*-acrylamidated cysteine residues were detected in the digest of archived BSA. Thus we concluded that proteins embedded in a polyacrylamide matrix largely remained intact even after years of storage.

Taken together, the results of experiments with BSA protein standard demonstrated that neither drying of gels nor their long-time storage noticeably affected the pattern of peptide mass maps and the recovery of individual peptides.

Table 5.

Quantification of the recovery of BSA tryptic peptides from archived and from fresh gels

Reak No	*m/z*	Sequence	$R_c$wet/R_c dried (%)*
I	927.49	YLYEIAR	103
II	1305.71	HLVDEPQNLIK	104
III	1439.81	RHPEYAVSVLLR	100
IV	1479.79	LGEYGFQNALIVR	104
V	1567.74	DAFLGSFLYEYSR	83
VI	1639.93	KVPQVSTPTLVEVSR	100

* The ratio of the relative concentration (Rc) of the peptide in the *in-gel* digest of the band excised from the wet gel to the Rc of the same peptide in the digest of the band excised from the archived gel. To determine Rc, five MALDI spectra were acquired from each of the samples and the results were averaged.

Table 6.

Peptides detected by MALDI-MS in the digests of BSA from the conventional gel and from the archived gel (spot 4). Detected peptides are designated with filled circles. Abbreviations are the same as in Table 7.

m/z	Residues	Sequence	Conventional gel	Archived gel
847.50	242 - 248	LSQKFPK	•	
927.49	161 - 167	YLYEIAR	•	•
1083.59	161 - 168	YLYEIARR	•	•
1142.71	548 - 557	KQTALVELLK	•	
1163.63	66 - 75	LVNELTEFAK	•	•
1193.60	25 - 34	DTHKSEIAHR	•	•
1249.62	35 - 44	FKDLGEEHFK	•	•
1283.71	361 - 371	HPEYAVSVLLR	•	
1305.71	402 - 412	HLVDEPQNLIK	•	•
1415.68	569-580	TVM(ox)ENFVAFVDK		•
1433.70	89 - 100	SLHTLFGDELC(acr)K		•
1439.81	360 - 371	RHPEYAVSVLLR	•	•
1479.79	421 - 433	LGEYGFQNALIVR	•	•
1553.83	483 - 495	LC(acr)VLHEKTPVSEK		•
1567.74	347 - 359	DAFLGSFLYEYSR	•	•
1639.93	437 - 451	KVPQVSTPTLVEVSR	•	•
1754.84	469 - 482	M(ox)PC(acr)TEDYLSLILNR		•
1809.84	387 - 401	DDPHAC(acr)YSTVFDKLK		•
1842.86	372 - 386	LAKEYEATLEEC(acr)C(acr)AK		•
1894.93	508 - 523	RPC(acr)FSALTPDETYVPK		•
2034.06	588 - 607	EACFAVEGPKLVVSTQTALA		•
2045.03	168 - 183	RHPYFYAPELLYYANK		•
2555.18	118 - 138	QEPERNEC(acr) FLSHKDDSPDLPK		•

3.1.3.3 Identification of proteins from archived gels

Gel archiving and long-time storage did not influence the number and the yield of BSA tryptic fragments (section 3.1.3.2). To further support this notion, we examined patterns of tryptic peptides yielded by *in-gel* digestion of a variety of protein spots with different *pI* and MW, excised from archived gels.

Digests of archived proteins were analyzed first by MALDI peptide mass mapping, and then, if required, by nanoES-MS/MS or MALDI QqTOF MS in line with the strategy outlined in the section 3.1.2. A combination of these methods enabled to establish the identity of a majority (> 97 %) of prominent peptide peaks observed in the spectra of digests. All spots from the two-dimensional gel (Figure 16A) were successfully identified and the overview of the results is presented in Table7.

Molecular weights of identified proteins were from 25 to 150 kDa and their *pI* also varied broadly. Most of protein sequences were submitted to a database upon completion of a human genome [142, 143], therefore *de novo* sequencing and cloning would have been required to characterize 9 out of 14 spots from the gel at the time when the proteins were isolated. Today, the availability of full-length sequences enabled to identify these proteins directly by MALDI MS, thus underscoring the importance of maintaining organized gel archives.

Although the sequence of a chicken protein (band p80 Fig. 16B) from a gel archived 15 years ago is still unknown, larger organismal representation in sequence databases and recent advances in mass spectrometry-driven sequence similarity searches [144] allowed us to identify it. Using a combination of nanoES MS/MS sequencing with a sequence-similarity searching tool MS BLAST [59] the chicken protein was confidently matched to its known human homologue.

Table 7
Identification of 14 spots from archived 2-D gel

Spot no	Protein	Acc. number, submission date	Peptide masses, matched / unmatched [a]	Sequence coverage, %	Met residues[b] Σ / M(ox) / M&M(ox)	Cys residues[c] Σ / C(acr) / C&C(acr)	Trp residues[d] Σ / W(ox) / W&W(ox)
1	SWI/SNF complex 155 kDa SWI/SNF complex 170 kDa	Q92923;Q92922 from 02.97	32 / 2[e]	25 / 12	13 / 4 / 1[e]	6 / 4 / 1[e]	4 / 2 / 0[e]
2	Fibronectin receptor CD29	P05556 from 11.88	9 / 1	12	2 / 2 / 0	2 / 0 / 1	___
3	DEAD box protein 1	Q92499 from 02.97	37 / 0	52	6 / 5 / 1	8 / 2 / 6	3 / 1 / 1
4	Bovine serum albumin	P02769 from 07.86	20 / 0	42	2 / 2 / 0	9 / 8/ 0	___
5	Heat shock 71 kDa protein	P11142 from 07.89	21 / 0	37	7 / 4 / 2	1 / 0/ 0	1 / 1 / 0
6	Hypothetical 55.2 kDa protein	Q9Y3I0 from11.99	15 / 0	30	6 / 5 / 0	1 / 0 /1	3 / 1 / 1
7	Ribonucleoprotein H	P31943 from 07.93	19 / 2	45	4 / 2 / 2	3 / 2 / 0	1 / 0 / 0
8	Actin-related protein BAF53A	O96019 from 05.99	16 / 1	40	4 / 1 / 1	1 /1 / 0	2 / 0 / 0
9	BAF57 protein	O43539 from 06.98	29 / 1[e]	55	8 / 1 / 0	___	4 / 2 / 1[e]
	TGF protein[f]	Q92734 from 02.97		10	2 / 2 / 0	___	
10	Disulfide isomerase	Q99778 from 05.97	17 / 1	50	___	1 / 1 / 0	2 / 0 / 1
11	γ-Actin	P02571 from 07.86	11/ 0	45	4 / 4 / 0	1 / 1 / 0	3 / 1 / 0
12	Ribonucleoprotein	Q9Y4J5 from 11.99	9 / 2	41	5 / 5 / 0	___	___
13	Transducin β chain 1	P04901 from 08.87	16 / 0[e]	35	1 / 0 /1[e]	6 / 6 / 0[e]	2 / 0 / 2[e]
	Transducin β chain 2	P11016 from 07.89		38			
14	Homebox Prox 1 protein	Q9Y224 from 11.99	21 / 1	65	1 / 0 / 0	2 / 0 / 2	3 / 1 / 1

a) Number of unmatched peaks having the intensity over 25% of the intensity of the most abundant peak in the spectrum
b) Methionine residues in detected peptides. Σ, total number of methionine residues; M(ox), detected in sulfoxide form only; M&M(ox), detected both in native and sulfoxide forms
c) Cysteine residues in detected peptides. Σ, total number of cysteine residues; C(acr), detected in *S*-acrylamide form only; M&M(ox), detected both in *S* -acetamide and Σ-acrylamide forms
d) Tryptophane residues in detected peptides. Σ, total number of tryptophane residues; W(ox), detected only in oxidized forms; W&W(ox), detected both in native und oxidized forms
e) Calculated in total for both identified proteins
f) Protein identified by nanoES-MS/MS

3.1.3.4 Modifications of archived proteins and confidence of database searches

By altering masses of amino acid residues modifications affect the confidence of protein identification. In this section we attempted to account the protein modifications specifically induced by archiving and storage of gels, and to identify the optimal method of considering their characteristic pattern in database searches.

First, MALDI peptide maps acquired from archived proteins (see the section above) were examined in detail and a second pass searching algorithm [52] was employed to account for common protein modifications, such as oxidation of methionine residues and acrylamidation of cysteine residues. If intense peptide ions were observed that could not be matched to the sequence of the identified protein or could not be explained by common protein modifications, the recovered pool of tryptic peptides was further analyzed by complementary mass spectrometric techniques. Acquired MS/MS spectra were searched against a database first in conventional ”stringent” mode, and then, upon manual interpretation, in error-tolerant mode using either partially matching sequence tags [57, 58] or MS BLAST sequence-similarity searches [59, 119].

Three major types of protein modification were detected in the analysis of protein spots from archived gels: oxidation of methionine residues, acrylamidation of cysteine residues and double oxidation of tryptophane residues (Table 7). These modifications are also frequently encountered in sequencing of proteins from conventional polyacrylamide gels [145, 146] and they are not caused specifically by gel archiving and storage. However, abundant peptides containing monooxidized methionine residues or dioxidized tryptophane residues were observed even though unmodified forms of these peptides were barely detectable, contrary to the modification patterns usually observed in peptide mass maps of proteins from conventional gels [147].

In addition to common protein modifications, comprehensive mass spectrometric analysis of the spot 13 revealed acetylation of the N-terminal amino acid residue of the respective protein. This protein modification also frequently occurs in eukaryotic cells. Importantly, no evidence of other modifications that could be attributed to oxidative damage of archived proteins, or to unspecific cleavage of these proteins by some exogenous proteases, was obtained.

Considering these modifications in database searches increases the statistical confidence of hits, especially if proteins were analyzed by MALDI peptide mass mapping. For example, a peptide mass map was acquired from an *in-gel* digest of a 38 kDa protein from spot 12 (Fig.16A), and masses of 14 peptide peaks were used for searching a database with a mass tolerance better than 100 ppm (Figure 17). The search produced no unequivocal identification. Database search was then repeated with the same set of peptide masses, but assuming that methionine residues are present only in the oxidized form (methionine sulfoxides) and cysteine residues are present only in acrylamidated form (cysteine *S*-acrylamide). The 37 kDa human Q9Y4J5 ribonucleoprotein then appeared at the top of the list with nine matching peptides, covering more than 50% of the protein sequence. Only two peptide ions of low abundance were unmatched and no human proteins other than Q9Y4J5 appeared in the list of hits.

Thus accurate accounting of protein modifications together with the quantitative estimation of the yield of *in-gel* digestion peptide products indicated that archiving of gels (especially, by drying them between sheets of a cellophane film) is a safe method of preserving proteins for the future analysis. Observed patterns of tryptic peptides and their recovery remained almost unchanged, compared to the ones from freshly prepared gels. The rarely observed differences were confidently attributed to common and well-understood protein modifications (such as oxidation of methionines and tryptophanes, or acrylamidation of cysteines), which, however, were “enhanced” by gel archiving.

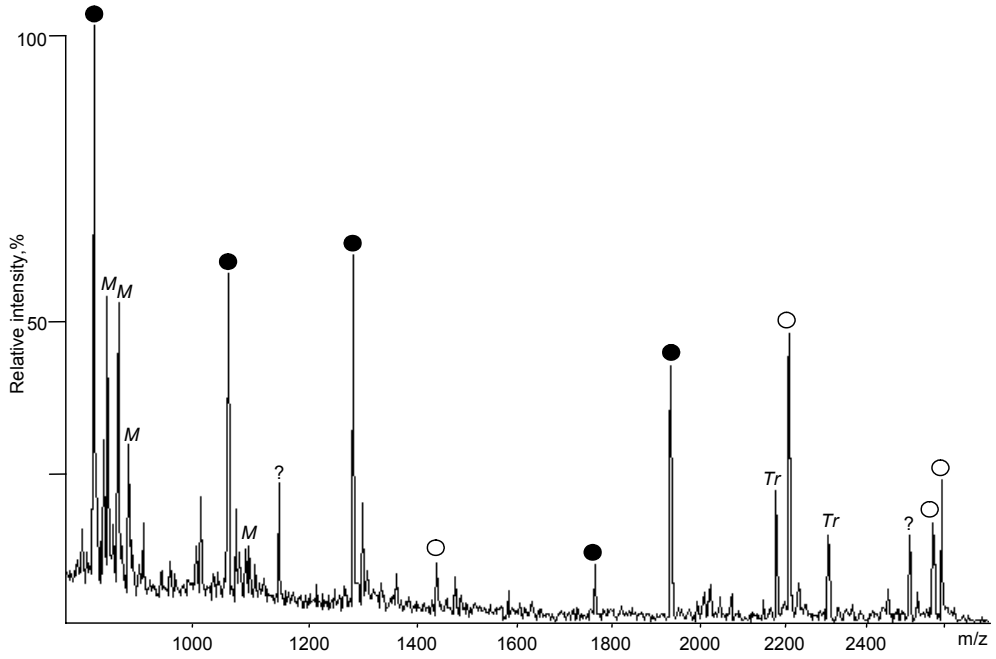


Figure 17. Identification of the spot 12 (see gel pattern in Figure 16A) by MALDI peptide mass mapping. Five peptides (peaks labeled with filled circles) were matched to the sequence of the Q9Y4J5 protein when the mass of native methionine was used in database searching. However, additional four peptides were matched (peaks labeled with unfilled circles) when methionine masses were set to the mass of methionine sulfoxide thus producing a very strong hit. Trypsin autolysis products are designated with Tr. M, matrix peaks. Two peptide peaks (labeled with a question mark) were not matched [134].

3.2. Analytical aspects of Tandem Affinity Purification (TAP) method in the budding and fission yeasts

Immunoaffinity purification followed by mass spectrometry is one of the most efficient proteomic approaches for dissecting protein complexes and protein interaction networks. A double-tag Tandem Affinity Purification method (termed TAP [89]) was developed and applied in a number of functional proteomics projects in *S. cerevisiae* and *S. pombe*. The overview of the TAP method from the perspective of mass spectrometric identification and characterization of subunits of protein complexes is presented in this chapter and is meant to address several important questions:

- whether TAP approach offers higher purification efficiency and better reproducibility of patterns of isolated interaction partners, compared to other affinity purification methods;
- how the success of TAP isolations depends on physicochemical properties of bait proteins;
- what is typical protein background of TAP purifications, is it reproducible and how can it be accounted for in order to reduce the rate of false positive interactors;
- how TAP results correlate with well established genetic approaches, applied for the characterization of protein-protein interactions, such as yeast two-hybrid screening;

3.2.1 Pattern and recovery of associated proteins in TAP and non-TAP isolations

To determine if a two-step TAP method is more efficient in purifying native protein complexes than conventional single-step purification approaches, a *S. cerevisiae* protein was TAP-tagged and immunoisolations of its interaction partners (which are known to constitute a stable protein complex) were carried out using single and double step affinity chromatography, followed by their identification by mass spectrometry.

Snu71p (ORF name YGR013w), a known member of U1 snRNP with TAP-tag fused to its C-terminus was used as the bait for this experiment. U1 snRNP, a previously characterized multi-subunit protein complex [150, 151] was isolated from two liters of the budding yeast cell culture by the group of Dr. B. Seraphin (EMBL, Heidelberg). Co-purified proteins were separated by one-dimensional SDS PAGE and protein bands visualized by Coomassie staining were excised, *in-gel* digested with trypsin and identified by MALDI peptide mapping. Identified proteins were compared with proteins recovered by one-step purifications, which employed either CBP- or ProtA-affinity tags followed by TEV protease-mediated release of the complex. While several yeast U1 snRNP subunits were recovered after a one step CBP affinity purification, the protein pattern was obscured by a large number of contaminating proteins (Figure 18, lanes 1 and 2). A similar observation was made for the affinity purification using IgG beads, followed by TEV protease-mediated release (Figure 18, lanes 3 and 4). In this case, however, contaminants corresponded to abundant background proteins and TEV protease (lane 6). A relatively large amount of TEV protease was required to increase the yield of the complex because of the low efficiency of heterophase proteolytic cleavage performed on the beads surface.

In contrast, almost no background proteins were detectable in the TAP purified fractions (lanes 7 and 8).

In addition to known subunits, Prp42 protein (ORF name YDR235w), previously identified only in a strain lacking Nam8p, another member of U1 snRNP [150], was detected as a strong band in the TAP-purified complex. Furthermore, one of the proteins, Snu30p (ORF name YDL087c), was not previously known as a component of the yeast U1 snRNP. Purification of the yeast U1 snRNP with a TAP-tagged Snu30p confirmed that Snu30p is a *bona fide* subunit of this complex [89].

Thus we concluded that both purification steps of the TAP procedure are required for highly specific purification with very low background of unspecifically associated proteins.

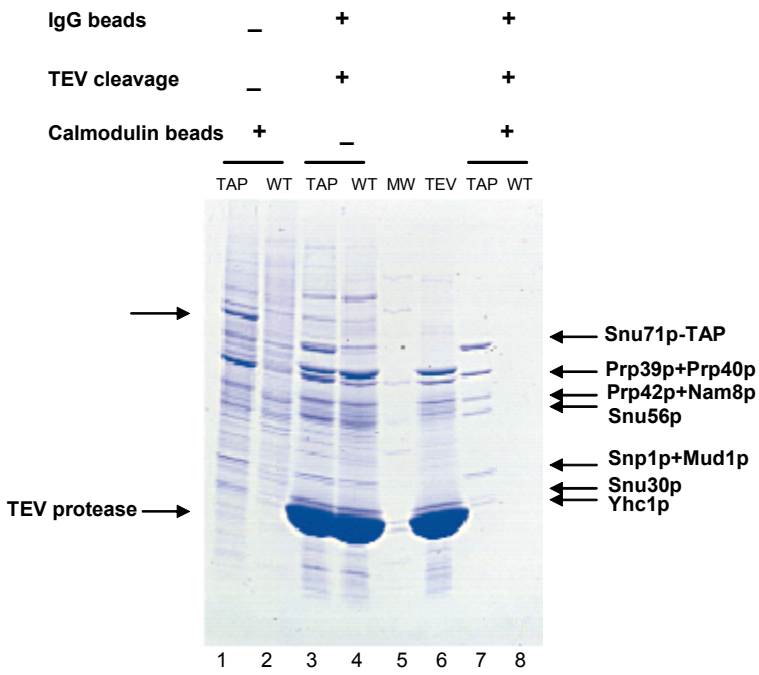


Figure 18. Analysis of protein profiles at various steps of the TAP purification of U1 snRNP complex. Proteins present in the final TAP fraction (lanes 7 and 8), or present after each of the single affinity purification steps (lanes 1–4), were identified by mass spectrometry. Snu71-TAP (lanes 1, 3, and 7) or wild-type extracts (lanes 2, 4, and 8) were used. Lane 5: molecular weight markers. Lane 6: an amount of TEV protease identical to the amount used to elute proteins bound to IgG beads (lanes 2, 3, 7, and 8). Arrows at the right-hand side indicate the U1 snRNP-specific proteins, including the bait Snu71-TAP; arrows at the left-hand side indicate bands of the Snu71p protein fused to the TAP tag and of the TEV protease [89].

We next asked if the TAP method isolates the same set of interactors as were observed in the direct immunoprecipitation using antibodies raised against the bait protein. To this end *S. cerevisiae* protein Esa1p (ORF name YOR224w), a member of NuA4 histone acetyltransferase protein complex [152] was isolated by TAP-method and its interactors were separated by SDS PAGE and identified by MALDI peptide mass fingerprinting (Figure 19A). The pattern and efficiency of TAP IP was compared with the purification by anti-Esa1p monoclonal antibodies reported previously by Galarneau *et al.* [153] (Figure 19B). Patterns of identified interactors looked very similar on both gels. However, proteins co-purified by anti-Esa1p antibodies could be visualized only by silver staining and only seven out of 11 known

subunits of the acetyltransferase complex could be identified by mass spectrometry. At the same time, TAP purification of the NuA4 complex rendered individual subunits detectable as intense Coomassie-stained bands, whose identification by MALDI peptide mapping was straightforward. Apart from seven known Esa1p interactors, five novel subunits were discovered. To confirm that identified proteins are *bona fide* members of the NuA4 complex, they were also TAP-tagged and IPed (the successive tagging strategy is discussed below in the Section 3) and in all experiments several subunits of NuA4 were identified by mass spectrometry.

Thus we concluded that the TAP approach substantially outperforms conventional immunoisolation methods in respect to the recovery of specifically interacting proteins and is accompanied by remarkably low protein background.

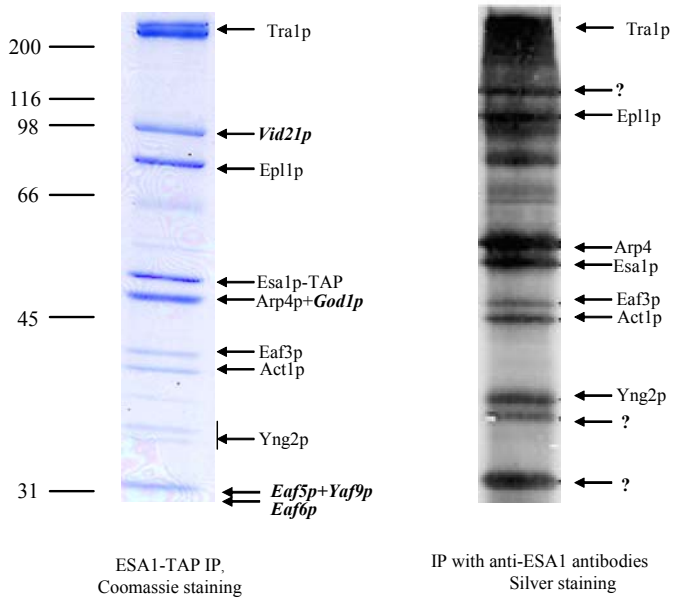


Figure 19. Isolation of Esa1p interactors by different immunoprecipitation methods. Panel at the left hand side: Coomassie-stained gel containing proteins immunoisolated by the TAP method with Esa1-TAP as bait. Gene names in italic indicate five novel protein subunits [this study].

Panel at the right hand side: Silver-stained gel with proteins purified by anti-Esa1p antibodies (the image was reported by Calarneau *et al*. in [153]).

Table 10.

Protein complexes from *S. cerevisiae* and *S. pombe* characterized by TAP MS approach

N	Protein complex	New complex	New members	Entry points	Cellular process	Reference
S.cerevisiae						
1	Set1C	yes	8	8	histone methylation	[137]
2	DNA-directed RNA polymerase II		1	2	transcription	[154]
3	Set3C	yes	7	6	histone deacetylation	[155]
4	Sum1C	yes	3	2	chromatin remodeling	[155]
5	NuA4		5	5	histone acetylation	*in progress*
6	Snt2C	yes	3	2	transcription	*in progress*
7	CPF		1	1	mRNA processing	[137]
8	Swr1C	yes	13	7	chromatin remodeling	*in progress*
9	Ino80C			1	chromatin remodeling	*in progress*
10	Rpd3C		7	6	histone deacetylation	*in progress*
11	IOC			1	budding	
12	Protein phosphatase 2A		1	1	cell cycle	
13	APC		2	3	cell cycle	*in progress*
14	Hda1C			1	histone deacetylation	
15	U1snRNP		1	1	splicing	[89]
16	U2snRNP		1	2	splicing	[156]
17	U4/U5/U6		3	3	splicing	[157]
18	Mak31/10/3	yes	3	1	RNA metabolism	[89]
19	CBC		1	1	RNA capping	[89]
20	RENT			2	chromatin silencing	[158]
21	Reb1C		4	1	transcription	*in progress*
S.pombe						
22	Sp_Set1C	yes	8	7	histone methylation	[122]
23	Sp_Lid2C	yes	5	4	histone methylation	[122]
24	Sp_Rad18C	yes	4	2	ubiquitination	*in progress*
25	Sp_CPF	yes	13	2	RNA processing	[123]

3.2.2 Protein background in TAP isolations in *S. cerevisiae* and *S. pombe*

Conventional IP experiments often result in complex mixtures of co-isolated proteins [159]. Because of the two affinity purification steps, isolations by the TAP method demonstrated much "cleaner" background, compared to conventional one-step affinity purifications of tagged proteins, although the method does pull down a few contaminants. Furthermore, scaling up does not improve the ratio between true interactors and background proteins and the latter are also enriched. It was therefore necessary to characterize the TAP protein background in detail in order to reliably distinguish *bona fide* interactors and minimize the risk of fetching false positives.

Usually, IPed proteins are separated on a one-dimensional polyacrylamide gel, and the pattern of proteins observed in the experiment lane is compared with the one in the control lane so that only proteins specifically detected in the experiment are subjected to further investigation by mass spectrometry. However, this approach is inherently error-prone. In our experiments all bands detected in the experiment lane were excised and analyzed by mass spectrometry - effectively, no control was used. On average, TAP isolation pulled down 30-60 Coomassie stainable bands of varying intensity and purity. The gel image of proteins co-purified with Swd2.1-TAP bait in *S. pombe* serves as a representative example. In total, 52 bands were excised, *in-gel* digested, analyzed by MALDI peptide mapping from this IP (Fig. 20) and 69 proteins were confidently identified (Table 8).

A list of proteins that were detected repeatedly in independent IPs with various TAP-tagged baits in the budding and fission yeasts is presented in Table 9 and was always subtracted from the list of potential interactors.

These proteins vary in the function, molecular weight and *pI*. We employed Codon Adaptation Index (CAI) to characterize the relative expression of the corresponding genes. CAI index uses a reference set of highly expressed genes from various species to assess the relative merit of each codon, and a score for a gene is calculated from the frequency of usage of all codons in that gene [161]. CAI of background proteins listed in Table 9 were, on average, 0.65 (*S. cerevisiae*) and 0.6 (*S. pombe*), thus indicating that these proteins were highly expressed.

In addition to these common proteins, a few contaminants were occasionally detected in some series of IPs, most likely because of minor variations in the quality of reagents, in particular of calmodulin beads.

Protein background in *S. pombe* (Table 9B) is different from *S. cerevisiae* (Table 9A), although, once again, highly abundant proteins, including housekeeping proteins, heat shock proteins and components of protein synthesis machinery, were typically observed.

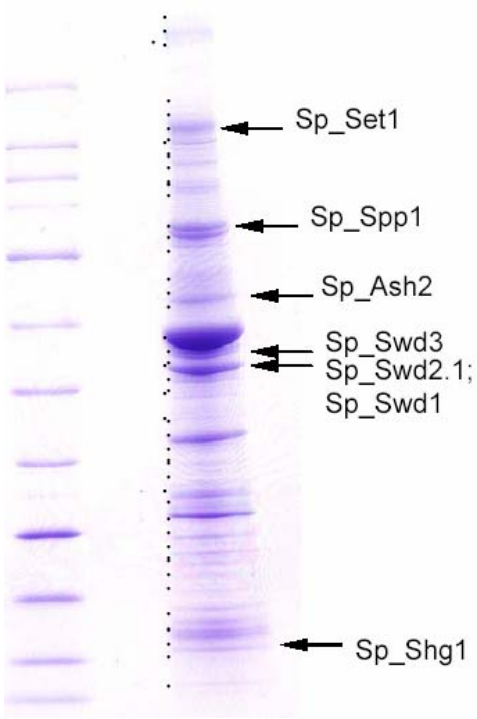


Figure 20. *S. pombe* Swd2.1/SPBC18h10.06c TAP immunoprecipitation. Left hand side lane: molecular weight markers (212, 158, 116, 97, 66, 55, 42, 36, 26, 20, 14 and 6 kDa). Right hand side lane: Tandem affinity purification using Swd2.1-TAP as a bait. 52 bands, which were excised and analyzed by mass spectrometry, are indicated by dots at the left hand side from the experimental lane. Bands of true Sp_Swd2.1 interactors are designated by arrows [123].

Table 8.

Proteins identified in TAP IP of *S. pombe* Swd2.1 (ORF name SPBC18h10.06c).

Band N	Protein name	Gene name	MW, kDa	Matched peptides	Sequence coverage
1	*too little*				
2	Fatty acid synthase, subunit beta	*FAS1*	230	12	6%
	Acetyl-CoA carboxylase	*CUT6*	256	11	5%
3,4	Putative glutamate synthase	*SPAPB1E7.07*	232	17	9%
5	*too little*				
6	Probable transcription initiation protein	*SPT5*	108	15	16%
7,8	Transcription silencing protein	*SET1*	105	19	20%
9	Eukaryotic translation initiation factor 3	*TIF32*	107	7	8%
10	Sulfite reductase [NADPH] flavoprotein	*SPCC584.01C*	111	19	17%
11	Coatomer gamma subunit	*SEC21*	101	11	16%
12	6-phosphofructokinase beta subunit	*PFK1*	103	16	17%
13	Hypothetical protein	*SPBC16H5.12c*	77	14	19%
14	Translation initiation factor eIF-3 beta	*EIF3b*	84	10	13%
15	pat1p homolog	*SPBC19G7.10c*	83	7	11%
16	Ash2-trithorax family protein	*ASH2*	74	18	27%
17	Heat shock protein	*HSP70*	70	21	34%
18	DnaK-type molecular chaperone	*SKS2*	67	12	21%
19	Acetolactate synthase, mitochondrial	*ILV1*	73	7	10%
20	Heat shock protein	*HSP60*	60	9	17%
21	Transcription silencing protein	*SET1*	103	19	17%
	T-complex protein 1	*CCT1*	60	8	17%
22	Phd finger transcription regulator	*SPP1*	49	9	20%
23,26	Elongation factor 1-alpha	*EF1-a*	50	20	43%
24	WD repeat protein	*SWD3*	42	10	27%
25	WD repeat protein	*SWD1*	45	10	23%
	WD repeat protein, the BAIT	*SWD2.1*	40	5	18%
27	Actin	*ACT1*	41	9	21%
	Ribosomal proteins	*RPL3A/B*	43		
28	Actin and elongation factor 1-alpha	*ACT1 and EF1-a*	41,50		
29	Ribosomal proteins	*RPL4A/B*	39		
30,31	as #23				
32	Glyceraldehyde-3-phosphate dehydrogenase	*TDH1*	36	14	29%
33,34	Slt1 protein	*SLT1*	49	6	16%
	Translation initiation factor eIF-3	*TIF35*	31	5	19%
35	Ribosomal proteins	*RPL5-1/2*	33		
36	26s proteasome regulatory subunit	*SPBC4C3.07*	33	7	26%
37	Ribosomal protein	*RPS3*	27		
38	Ribosomal proteins	*RPL13, RPL2-1/2, RPS1*	24/28		
39	Ribosomal proteins	*RPL7C, RPS4-1/2*	29		
40	Transcription silencing protein, N-termini	*SET1*	*103*		
41	Ribosomal proteins	*RPL1A/B*	25		
42	Ribosomal protein	*RPL15B*	24		
43	Ribosomal protein	*RPS7*	22		
44	Ribosomal proteins	*RPL17A/B*	21		
45	Ribosomal proteins	*RPL20A/B*	20		
46	Ribosomal proteins	*RPL11_1/2, RPL15*	20,24		
47	Ribosomal proteins	*RPS11A/B, RPL36*	10,19		
48	Ribosomal proteins	*RPS17-1/2, RPS13*	15		
49	Ribosomal proteins	*RPS10-1/2, RPS24A/B*	15		
50	Ribosomal proteins	*RPL35, RPS19A, RPS20*	15		
51	Hypothetical protein	*SHG1*	15	16	53%
52	Ribosomal proteins	*RPS25A/B*	10		

True Sp_Swd1.2 interactors confirmed by further experiments [122, 123] are highlighted.

Table 9.

Protein background in TAP IPs.

A. List of typical protein contaminants in *S. cerevisiae* TAP purifications

Protein name	Gene name	MW, kDa	*pI*	CAI[a]	Localization
Heat shock protein homolog	*SSE1*	77	5.0	0.521	Cytoplasm
Major coat protein of La virus	*GAG*	76	5.5		Cytoplasm
Heat shock proteins	*SSA2 / 1*	69.4	4.9	0.802/0.709	Cytoplasm
Heat shock proteins	*SSB2 / 1*	66.5	5.2	0.820/0.772	Cytoplasm
Pyruvate kinase 1	*CDC19*	54.1	7.7	0.893	Cytoplasm
Elongation factor 1-alpha	*TEF1*	49.9	9.2	0.871	Cytoplasm
Tryptophanyl-trna synthetase	*WRS1*	49.2	6.6	0.286	Cytoplasm
Enolase 2	*ENO2*	46.7	5.8	0.892	Cytoplasm
Phosphoglycerate kinase	*PGK1*	44.6	7.1	0.815	Cytoplasm
Glyceraldehyde 3-phosphate dehydrogenase 3	*TDH3*	35.6	6.7	0.924	Cytoplasm
Vacuolar ATP synthase subunit e	*VMA4*	26.3	5.2	0.263	Cytoplasm
Phosphoglycerate mutase	*GPM1*	27.6	8.9	0.811	Cytoplasm
60S ribosomal protein l3	*MAK8*	43.6	10.4	0.830	Cytoplasm
Putative ATP-dependent RNA helicase [b]	*DED1* [b]	65.4	7.9	0.376	Cytoplasm
Heat shock protein 60 [b]	*MIF4* [b]	58.2	4.9	0.382	Mitochondria
Mitochondrial import membrane translocase [b]	*TIM44* [b]	48.8	9.6	0.140	Mitochondria
Other ribosomal proteins[c]		10-30			
MEAN				0.65+/-0.23	

[a] for the method of calculation of Codon Adaption Index (CAI) see [161].

[b] proteins only observed in some immunoprecipitations

[c] background low molecular weight ribosomal proteins vary, with genes RPL32, RPS13, RPS15, RPS26A/B, RPL19A/B, RPS24A/B, RPS19A/B, RPS18A/B, RPL17A/B, RPL7A/B, RPL4A/B, RPS11A/B, RPS14A/B, RPL33A/B, RPL10, RPS7A/B, RPS6A/B, RPL2A/B being most typical.

B. List of typical protein contaminants in *S. pombe* TAP purifications

Protein name	Gene name	MW, kDa	*pI*	CAI*	Localization
Probable scetyl-CoA carboxylase	*CUT6*	257	6.2	0.373	Cytoplasm
Putative glutamate synthase	*SPAPB1E7.07*	233	6.1	0.516	Cytoplasm
Fatty acid synthase, subunit beta	*FAS1*	230	6.1	0.427	Cytoplasm
Fatty acid synthase, subunit alpha	*FAS2*	202	6.0	0.395	Cytoplasm
Translation initiation factor eIF-3 p110 subunit	*TIF32*	107	9.1	0.378	Cytoplasm
6-phosphofructokinase beta subunit	*PFK1*	103	6.0	0.710	Cytoplasm
Coatomer gamma subunit	*SEC21*	101	5.1	0.297	
Protein of unknown function	*SPBC16H5.12c*	77	5.8	0.278	?
Translation initiation factor eIF-3 subunit	*ELF-a*	50	9.3	0.879	Cytoplasm
Heat shock protein 70kDa family	*HSP70*	70	5.1	0.789	
Probable ATP-dependent RNA helicase	*DED1*	70	8.8	0.523	Cytoplasm
Heat shock protein 70kDa, mitochondrial	*SSC1*	73	7.0	0.601	Mitochondria
Heat shock protein 70kDa family	*SKS2*	67	5.9	0.802	Cytoplasm
Translation initiation factor eIF-1 alpha	*EFL-a*	50	9.3	0.879	Cytoplasm
Mitochondrial import inner membrane translocase	*SPBC14CH3.01*	49	9.5	0.257	Mitochondria
Probable DNA-J-like protein	*SPAC4H3.01*	45	6.3	0.237	Cytoplasm
DNA-J protein homolog	*SPJ1*	42	8.5	0.319	Cytoplasm
Actin	*ACT1*	42	5.3	0.719	Cytoplasm
Glyceraldehyde-3-phosphate dehydrogenase	*TDH1*	36	6.5	0.849	Cytoplasm
RVS161 protein homolog	*HOB3*	30	6.7	0.301	
Phosphoglycerate mutase	*GPM1*	24	7.2	0.849	Cytoplasm
Other proteins**		6-30			
MEAN				0.6+/-0.33	

* for the method of calculation of Codon Adaption Index see [161]

** Background ribosomal proteins may vary. The proteins Rpl3-1/2, Rpl2, Rpl5-1/2, Rpl13, Rps3, Rps1-1/2, Rps6, Rps7A/C, Rps9A/B, Rps11A, Rps13, Rps17-1/2, Rps22A, Rps18, Rps8, Rps25A /B, Rpl15-2, Rpl15, Rpl17, Rpl20-1/2, Rpl21-1/2, Rpl11A/B, Rpl24, Rpl28A/B, Rpl25A/B and Rpl36A/B were most typical contaminants.

3.2.3 Application of the TAP method for isolating protein complexes from the budding and fission yeasts

The TAP procedure followed by the identification of proteins by mass spectrometry was applied to characterize a large number of the intact protein complexes in *S.cerevisiae* and *S.pombe* (Table 10). The identified proteins represent a reliable dataset, which allowed us to estimate the success rate of TAP tagging and purification and to compare the identified interactors with the ones discovered by alternative methods (such as two-hybrid screening).

3.2.3.1 Success rate of TAP tagging and purification in the budding and fission yeasts

It is conceivable that physicochemical properties of bait proteins might determine the success of tagging and subsequent affinity isolation of interaction partners. However, it was not at all clear how the molecular weight, *pI*, abundance, post-transcriptional and post-translational modifications affect the overall success rate of the TAP method. Is there any difference between the success rate of tagging essential and non-essential genes? Is the success rate different in the two most popular model organisms within the fungi lineage, *S .cerevisiae* and *S. pombe*? These and other related issues have been addressed in this chapter by the critical evaluation of the results of the TAP method, applied to a variety of bait proteins in numerous collaborations projects with the laboratories of Drs. A.F. Stewart (EMBL / Technische Universität Dresden – BIOTEC) and B. Seraphin (EMBL, Heidelberg) (for projects overview and full list of references see [89, 123, 136]).

Altogether, the results of TAP tagging and immunoaffinity isolation of 75 baits from *S. cerevisiae* and 23 baits from *S. pombe* were evaluated. The complete list of attempted genes and TAP results is presented in Table 11A for the budding yeast and in Table 11B for the fission yeast.

For all *S. pombe* proteins and for the majority of *S. cerevisiae* proteins TAP-tag was introduced at their C-termini. C-terminal TAP modification is technically easier to perform and is thought to cause minimal disturbance of protein expression and folding. Although Puig *et al.* reported that in less than 5% of TAP-tagged proteins fusion of the tag to the C-terminus impaired protein function and led either to severe

growth defects or to cell death [120], in no case cell lethality or a severely disturbed phenotype was observed in this study. In a single case (TAP tag fused to the C-terminus of Set1 protein in *S. cerevisiae*) we observed that although the composition of the isolated Set1C complex was not perturbed, the tag impaired its catalytic activity [137]. Fusion of the tag to the N-terminus of Set1p allowed us to isolate the same complex having catalytic activity similar to the activity in the wild type. However, we urge to consider the achieved high success rate with caution, because of a relatively small selection of target genes.

The results of TAP-tagging and affinity isolation of 75 bait proteins in *S. cerevisiae* are presented as the pie-chart diagram (Figure 21). In three (4%) out of 75 attempted *S. cerevisiae* genes the bait was not detected by Western blotting, *i.e.* the tagging procedure failed or the tagged protein was not expressed. Four other bait proteins (5%) were only detected by Western blot analysis and it was not possible to identify baits by mass spectrometry because they were purified in too low amount.

TAP-tag was successfully fused with 72 proteins (96 %). Their molecular weights were between 9 and 175 kDa, calculated *pI* were between 4.5 and 10.0, and CAI from 0.071 to 0.614, indicating that TAP tagging was also successful for low expressed proteins. Among the 72 successfully tagged genes, 16 were essential, according to the YPD database [163] and no genes encoded for membrane proteins.

Altogether, products of 68 genes out of the total of 72 genes with successfully fused tag (91%) were recovered by immunoaffinity chromatography in amounts sufficient for their reliable identfication by mass spectrometry. Interaction partners were determined for 57 baits (76%). In 11 purifications (15%) only bait proteins were detected with no plausible interaction partners. These data strongly indicate that overall success of the TAP procedure in *S. cerevisiae* has not been strongly influenced by the size, *pI* and relative expression of target genes, or if the target gene was essential.

Table 11.

A. List of baits in *S. cerevisiae*

N	Gene name (ORF name)	TAP IP result[a]	*pI*/MW (kDa)	Null mutant	Chromosome	CAI[b]
1	***SET1*** (YHR119W)	+	9.16/124	Viable	VIII	0.146
2	***SET2*** (YJL168C)	+	8.79/84	Viable	X	0.138
3	***SET3*** (YKR029C)	+	8.99/85	Viable	XI	0.120
4	***SET4*** (YJL105w)	Western blot	8.98/64	Viable	X	0.102
5	***SET5*** (YHR207C)	Bait only	6.23/60	Viable	VIII	0.144
6	***SET6*** (YPL165C)	Negative Western blot	7.91/44	Viable	XVI	0.135
7	***HST1*** (YOL068C)	+	8.80/58	Viable	XV	0.133
8	***HST2*** (YPL015C)	+	5.63/40	Viable	XVI	0.142
9	***HST3*** (YOR025W)	+	9.38/50	Viable	XV	0.125
10	***HST4*** (YDR191W)	+	9.39/42	Viable	IV	0.123
11	***HOS1*** (YPR068C)	Bait only	5.16/55	Viable	XVI	0.116
12	***HOS2*** (YGL194C)	+	5.06/51	Viable	VII	0.123
13	***SIR2*** (YDL042C)	+	8.77/63	Viable	IV	0.148
14	***SIR3*** (YLR442C)	+	6.15/111	Viable	XII	0.122
15	***SIR4*** (YDR227W)	+	9.19/152	Viable	IV	0.136
16	***SNT1*** (YCR592)	+	9.30/138	Viable	III; Intron	0.120
17	***HOS4*** (YIL112W)	+	5.51/124	Viable	IX	0.139
18	***SIF2*** (YBR103W)	+	4.83/59	Viable	II	0.108
19	***CPH1*** (YDR155C)	Bait only	6.75/17	Viable	IV	0.614
20	***SUM1*** (YDR310C)	+	6.00/118	Viable	IV	0.134
21	***TOS4*** (YLR183C)	+	9.52/55	Viable	XII	0.148
22	***RFM1*** (YOR279C)	+	9.18/35	Viable	XV	0.127
23	***ESA1*** (YOR244W)	+	7.77/52	Lethal	XV	0.149
24	***SNT2*** (YGL131C)	+	8.90/163	Viable	VII	0.131
25	***BRE2*** (YLR015W)	+	5.89/58	Viable	XII	0.132
26	***RMS1*** (YDR257C)	+	4.86/57	Viable	IV	0.171
27	***SDC1*** (YDR469W)	+	4.61/19	Viable	IV	0.106
28	***SWD1*** (YAR003W)	+	4.64/49	Viable	I	0.129
29	***SWD3*** (YBR175W)	+	5.82/35	Viable	II	0.089
30	***SHG1*** (YBR258C)	+	4.90/16	Viable	II	0.144
31	***SWD2*** (YKL018W)	+	5.68/37	Lethal	XI	0.136
32	***SPP1*** (YPL138C)	+	6.58/41	Viable	XVI	0.134
33	***DOT6*** (YER088C)	Bait only	9.66/73	Viable	V	0.167
34	***REB1*** (YBR049C)	+	5.07/92	Lethal	II	0.199
35	***CHD1*** (YER164W)	Bait only	6.29/168	Viable	V	0.170
36	***NEW1*** (YPL226W)	Bait only	5.69/134	Viable	XVI	0.304
37	***RAD52*** (YML032C)	Bait only	8.63/56	Viable	XIII	0.132
38	***STU2*** (YLR045C)	Bait only	8.80/101	Lethal	XII	0.128
39	***YNG2*** (YHR090C)	+	8.33/32	Viable	VIII	0.177
40	***EPL1*** (YFL024C)	+	8.92/97	Lethal	VI	0.154
41	***GOD1*** (YGR002C)	+	9.50/55	Lethal	VII	0.121
42	***EAF3*** (YPR023C)	+	8.58/45	Viable	XVI	0.111
43	***YAF9*** (YNL107W)	+	5.19/26	Viable	IV	0.125
44	***ISW1*** (YBR245C)	+	6.42/131	Viable	II	0.192
45	***BAS1*** (YKR099W)	+	8.26/89	Viable	XI	0.115
46	***RPD3*** (YNL330C)	+	5.38/49	Viable	XIV	0.143
47	***SIN3*** (YOL004W)	+	5.45/175	Viable	XV	0.150
48	***PHO23*** (YNL097C)	+	7.83/37	Viable	XIV	0.133
49	***UME1 (***YPL139C)	+	5.09/51	Viable	XVI	0.133
50	***RCO1*** (YMR075w)	+	9.05/79	Viable	XIII	0.164
51	***SWR1*** (YDR334W)	+	6.03/174	Viable	IV	0.145
52	***SWC1*** (YAL011W)	+	9.26/72	Viable	I	0.146
53	***RVB1*** (YDR190C)	+	5.67/50	Lethal	IV	0.190
54	***ARP6*** (YLR085C)	+	5.37/50	Viable	XII	0.108
55	***HDA1*** (YNL021W)	+	5.34/80	Viable	XIV	0.151

56	***BDF1*** (YLR399C)	+	5.85/77	Viable	XII	0.127
57	***ZDS1*** (YMR273C)	+	6.14/103	Viable	XIII	0.131
58	***CDC16*** (YKL022C)	+	6.56/95	Lethal	XI	0.124
59	***SNU71*** (YGR013W)	+	4.83/71	Lethal	VII	0.134
60	***NAM8*** (YHR086W)	+	8.64/57	Viable	VIII	0.124
61	***LEA1*** (YPL213W)	+	7.38/27	Viable	XVI	0.109
62	***LSM8*** (YJR022W)	+	7.88/15	Lethal	X	0.134
63	***MUD13*** (YPL178W)	+	5.05/24	Viable	XVI	0.173
64	***RSE1*** (YML049C)	+	5.66/154	Lethal	XIII	0.132
65	***POP4*** (YBR257W)	+	9.42/33	Lethal	II	0.126
66	***DBP5*** (YOR046C)	+	7.28/54	Lethal	XV	0.211
67	***MEX67*** (YPL169C)	+	9.34/67	Lethal	XVI	0.120
68	***MAK31*** (YCR020C)	+	5.27/10	Viable	III	0.140
69	***LSM3*** (YLR438C-A	+	4.24/10	Lethal	XII	0.071
70	***KEM1*** (YGL173C)	+	7.29/175	Viable	VII	0.194
71	***ECM5*** (YMR176W)	Western blot	6.61/163	Viable	XIII	0.131
72	YBL054W	Western blot	9.51/59	Viable	II	0.142
73	YDR026C	Western blot	9.46/66	Viable	IV	0.155
74	***VID21*** (YDR359C)	No homologous recombination	9.00/109	Viable	IV	*not available*
75	***TRA1***(YHR099W)	No homologous recombination	6.32/433	Lethal	VIII	0.139

B. List of baits in *S. pombe*

N	Gene name	TAP IP result	*pI*/MW (kDa)	Null mutant	Chromosome	CAI[b]
1	SPCC306.04c	+	9.1/105	Viable	III	0.231
2	SPCC594.05c	+	6.0/49	Viable	III	0.204
3	SPBC13g1.08c	+	6.6/72	Viable	II	0.240
4	SPAC23h3.05c	+	5.2/45	Viable	I	0.234
5	SPBC18h10.06c	+	6.7/40	Viable	II, two introns	0.214
6	SPAC824.04	+	6.0/38	Viable	I	0.234
7	SPBC354.03	+	5.9/43	Viable	II	0.218
8	SPCC18.11c	+	4.9/12	Viable	III, one intron	0.223
9	SPAC17g8.09	Negative Western blot	5.1/15	Viable	I, two introns	0.177
10	SPAC17G6.16c	+	5.9/88	Lethal	I, five introns	0.250
11	SPBP19A11.06	+	8.1/172	?	II	0.219
12	SPAC3h1.12c	+	8.7/129	?	I	0.243
13	SPAC18B11.07c	+	6.3/17	Viable	I, four introns	0.267
14	SPAC22F8.12C	+	8.8/19	Viable	I	0.197
15	SPAC824.10c	+	5.0/37	Viable	I	0.240
16	SPAC22E12.11c	+	9.5/95	Viable	I	0.207
17	SPBC428.08c	Negative Western blot	8.7/56	Viable	II	0.238
18	SPBC16D10.07c	Bait only	5.0/53	Viable	II, five introns	0.239
19	SPBC1734.06	Bait only	8.1/43	Viable	II	0.219
20	SPAC27D7.05c	No homologous recombination	9.6/12	Viable	I	0.226
21	SPBC9B6.12c	+	5.1/16	Lethal	II	0.209
22	SPBC83.04	+	3.7/16	Viable	II, one intron	0.324
23	SPAC6F12.15c	+	5.6/75	Lethal	I, two introns	0.223

[a] "+" stands for TAP IP, in which both bait and interaction partners were pulled down in amounts sufficient for the identification by MS; "Bait only" stands for experiments, in which only bait was retrieved; "Negative Western blot" - bait was not detectable even by Western blot; "No homologous recombination" stands for experiments, in which fusion of TAP tagging cassette with corresponding genes failed; "Western blot" – bait was only detectable by Western, but not by MS.

[b]. CAI – Codon Adaption Index [161].

In total, 23 genes from *S. pombe* were TAP-tagged. Molecular weights of bait proteins were in the range of 12 to 172 kDa and their *pI* were from 5 to 9.5. Their codon adaptation indexes (CAI) were from 0.177 to 0.324 hence suggesting that the selected proteins were relatively low abundant. Three out of these 23 genes were essential. The TAP tag was successfully fused to the C-termini of 22 proteins, and 20 tagged proteins (87%) were subsequently detected by Western blot. Two bait proteins did not pull down any detectable interaction partners, although the baits themselves were visualized as intense Coomassie stained bands. Each of the other 18 baits (78%) pulled down 2-12 interaction partners.

Comparison of the overall success of TAP-tagging and isolation of interaction partners suggested that it was very similar in the budding and fission yeasts (91% and 87%, respectively). The number of "productive" baits that successfully pulled down detectable amounts of interaction partners was also very similar.

Proteome-wide purification of protein complexes by TAP approach in the budding yeast [1] reported that from 1739 processed genes 1167 (67%) expressed the tagged proteins at the detectable level and only 589 tagged proteins (34%) were successfully purified. The success rate achieved in this work was significantly higher: 96% and 81%, respectively. We attribute this disagreement mostly to the limitations imposed by the adopted high throughput strategy, since Gavin *et al.* [1] has attempted 23 times more genes than described here. Furthermore, assuming the budding yeast genome encodes for about 6,000 proteins [164, 165], the study by Gavin *et al.* attempted to cover 29 % of the budding yeast proteome and provided much less biased selection of baits. Interestingly, the set of 1739 genes reported by Gavin *et al.* included 293 genes encoding for membrane or membrane-associated proteins, and the success TAP tagging in this category (14%) was remarkably below the average.

Despite large difference in the reported success rates, this study of Gavin *et al.* suggested a similar ratio of "productive" baits (baits, which pulled down interaction partners) to the total number of baits purified at the mass spectrometry detectable level. TAP-isolation of 78% of 589 gene products enabled to identify plausible interaction partners [1] vs 83% (57 from 68 successfully tagged *S. cerevisiae* genes) reported in this work.

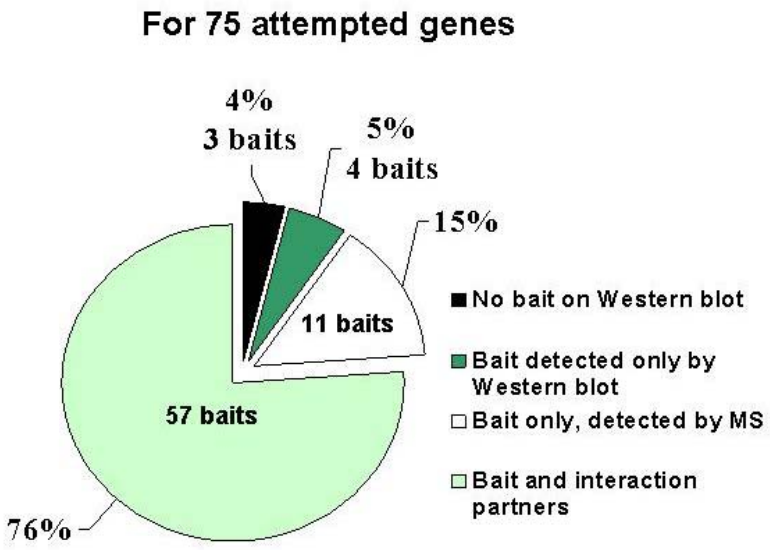


Figure 21. Pie-chart diagram summarizing the results of immunoprecipitation of TAP-tagged *S. cerevisiae* proteins.

3.2.3.2 Why TAP method sometimes fails?

We inspected in detail all cases in which either TAP-tagging of *S.cerevisiae* proteins failed or mass spectrometry did not identify the bait. To find out if other methods could suggest plausible interactors when TAP-tagging, for any reason, failed, data from genome-wide two-hybrid screens, bioinformatics predictions or previously reported genome-wide tagging studies [1, 2] accumulated in the knowledge databases [163, 166] were critically considered. The results are summarized in Table12.

Fusion of the TAP tag to *VID21* and *TRA1* failed most likely because their ORFs were predicted incorrectly. Both proteins were observed as intense Coomassie bands and confidently identified by MALDI peptide mapping in TAP IPs with other subunits of the corresponding protein complexes, suggesting that these proteins are expressed at the reasonably high level. The information obtained by other methods

(Table 12) also confirmed that *TRA1* and *VID21* are expressed, functional and participate in various protein-protein interactions.

In one case (*SET6*), homologous recombination was successful, but Western blot detected no fusion protein. The low Codon Bias Index (CBI) [160] and Codon Adaption Index (CAI) [161] (0.063 and 0.135 respectively) indicate that *SET6* is a low expressed protein and probably it was even less expressed in the mutant strain.

Four baits (*SET4*, *ECM5*, YBL054W and YDR036C) were detected only by Western blot. Fusing the TAP tag to N-termini of *SET4* and *ECM5* did not improve the yield, as well as attempts to purify them from a larger volume of cell culture. Although CAI index of Ecm5 protein is relatively low, it was identified as a specific interactor in two other TAP IPs, in which Rpd3 and Snt2 proteins were used as baits and was confirmed to be a *bona fide* member of *S. cerevisiae* Snt2C compex [122]. Protein-protein interactions of Ecm5p were also reported by others [1].

For the two (*SET4* and *SET6*) out of seven baits, in which TAP approach failed, other methods also did not suggest any plausible protein-protein interactions and their biological function remains unknown.

Table 12.

Genes whose TAP-tagging and/or IP failed.

Gene/ORF names	Tag position	Homologous recombination	Western blot	Protein properties	2HY[a]	Protein-protein interactors[b]	Complexes[bc]	Cellular Role
SET4 YJL105w	C, N	+	+	*pI* 8.98/64kDa Null: Viable CAI: 0.102		_	_	Unknown
*ECM5** YMR176W	C, N	+	+	*pI* 6.61/163kDa Null: Viable CAI: 0.131		_	One complex	Unknown
YBL054w	C	+	+	*pI* 9.51/59kDa Null: Viable CAI: 0.142	1	_	_	Unknown
YDR026C	C	+	+	*pI* 9.46/66kDa Null: Viable CAI: 0.155	4	_	_	Unknown
SET6 YPL165C	C	+	Not detected	*pI* 7.91/44kDa Null: Viable CAI: 0.135		_	_	Unknown
*VID21** YDR359C	C, N	Negative	-	*pI* 9.00/110kDa Null: Viable		_	Three complexes	Unknown
*TRA1** YHR099W	C	Negative	-	*pI* 6.32/443kDa Null: Lethal CAI: 0.139	1	23	11 complexes	Transcription factor

* proteins were identified as potential interactors in other IPs in this study; [a] interactors identified in high-throughput two hybrid studies; [b] source: YPD ; [c] from Gavin *et al.* [1] and Ho *et al*. [2].

3.2.3.3 Protein interactors identified by TAP - mass spectrometry approach

In *S. cerevisiae* 57 "productive" baits pulled down a total of 415 interaction partners, averaging 7.3 interactors per protein, which agrees well with independent bioinformatic estimates [103, 167]. The number of interactors purified with the bait varied between 1 and 21. The molecular weight of protein interactors was in the range of 8 to 430 kDa and their *pI* also varied broadly. Both low expressed proteins as well as highly abundant proteins were identified (note that background proteins from Table 9A were not considered here). Similar estimates were reported by Gavin *et al.* [1]. Thus we concluded that TAP method provides truly unbiased recovery of protein interactors, which does not notably depend on their physicochemical properties.

In 11 cases only baits were detected in TAP IPs (Table 13). Five of these idle baits (*HOS1, HST2, HST4, SET5* and *BAS1*) were not reported to participate in protein-protein interactions and were not assigned to any protein assemblies by other methods. These five proteins are of "average" molecular weight and *pI* although their CAI values are rather low. Fusing the TAP tag at the N-terminus of *SET5* did not help to identify protein interactors.

For other six baits (*NEW1, CPH1, DOT6, RAD52* and *STU2*) numerous protein-protein interactions were reported. Two of them, New1p and Cph1p, were also identified as specific proteins in purifications with several baits listed in Table 7. These two proteins are highly expressed (CAI 0.304 and /0.614 respectively) and were visualized as intense Coomassie stained bands on the gels with their TAP IPs. We therefore speculate that these two proteins are mostly present in a "free" form, whereas only a small fraction is actually involved in protein-protein interactions. Hence the unfavorable ratio between the amount of the bait and its interactors did not enable their identification by TAP MS approach.

Critical analysis of results of TAP purifications in both model organisms revealed several limitations of the method in its ability to provide a comprehensive overview of protein-protein interactions. First, unfavorable stoichiometric ratios between subunits of the complex might compromise the completeness of the characterization of its composition. Transient interactors that are only present in a substoichiometric amount in comparison with core subunits might be missed by mass spectrometric identification. Second, topography of the protein complex might only

allow purifying a fraction of its subunits, rather than the entire complex. Third, some proteins might be shared between more than one protein complex and a subset of subunits of several distinct protein complexes might be co-isolated, if such proteins were inadvertently chosen as baits.

Table 13

Idle baits in the TAP MS method

Gene/ ORF names	Protein propetries	Tag position	2HY[a]	Protein-protein interactions[b]	Complexes[bc]	Cellular role
HOS1 YPR068C	*pI* 5.16; 55kDa; Null: Viable CAI: 0.116	C	_	_	_	Histone deacetylation
HST2 YPL015C	*pI* 5.63; 40kDa; Null: Viable CAI: 0.142	C	_	_	_	Fatty acid catabolism; Chromatin silencing
HST4 YDR191W	*pI* 9.39; 42kDa; Null Viable CAI: 0.123	C	_	_	_	Short-chain fatty acid catabolism Chromatin silencing
SET5 YHR207C	*pI* 6.23; 60kDa; Null: Viable CAI: 0.144	C,N	2	_	_	Transcription
BAS1 YKR099W	*pI* 8.26; 89kDa; Null Viable, CAI: 0.115	C	5	_	_	Transcription, nucleoside metabolism
*NEW1** YPL226W	*pI* 5.69; 134kDa; Null: Viable CAI: 0.304	C	_	_	two complexs	Unknown
*CPH1** YDR155C	*pI* 6.75; 17kDa; Null: Viable CAI: 0.614	C	2	6	19 complexes	Protein-nucleus export, response to stress, histone deacetylation
DOT6 YER088C	*pI* 9.66; 73kDa; Null: Viable CAI: 0.167	C	1	_	One complex	Chromatin silencing
CHD1 YER164W	*pI* 6.29; 168kDa; Null: Viable CAI: 0.170	C	1	9	two complexs	Transcription Chromatin assembly/disassembly
RAD52 YML032C	*pI* 8.63; 56kDa; Null: Viable CAI: 0.132	C	5	2	4 complexes	Double-strand break repair
STU2 YLR045C	*pI* 8.80; 101kDa; Null:Lethal CAI: 0.128	C	4	5	_	Microtubule stabilization

*was found in other TAP IPs with baits listed in the Table 7 in this study; a – interactors identified in high-throughput two hybrid studies; b - from YPD, c – from genome-wide screens by Gavin *et al.* [1] and Ho *et al.* [2].

For better consistency, all TAP IPs in this work were carried out under the same experimental conditions. As was demonstrated previously [168], salt concentration in a buffer, which is used during TAP purification might strongly influence the observed pattern of protein interactors. By decreasing the NaCl concentration in calmodulin binding, washing and elution buffers from 150 to 100 mM Krogan *et al.* [168] additionally purified six proteins specifically binding the TAP tagged protein Chd1 (Table 13). Although purifying complexes under "mild" conditions enhances background, it also increases the chance to recover weakly bound interactors.

A variability of properties of "idle" protein baits (Table 13) strongly indicates that a variety of poorly understood technical and physicochemical factors might be responsible for these "unproductive" purifications.

3.2.3.4 Interactors identified by TAP MS *vs* interactors identified by yeast two-hybrid screening

Yeast two-hybrid screen (2HY) is a recognized tool for mapping of protein-protein interactions at the genome-wide scale [77, 169]. Considerable efforts have already been invested in comprehensive charting of the budding yeast "interactome" [91, 93, 94]. It was therefore important to understand how protein-protein interactions detected by 2HY correlate with interactors identified by TAP-MS approach.

We compared interactors identified by TAP-MS with interactors suggested for the same baits by 2HY (Figure 22). Since it is known that 2HY data obtained by different groups are in poor concordance and also might miss up to 90% of known interactions [91], we considered protein-protein interactions discovered in several studies and annotated in the Incyte knowledge database [163]. For the same set of 57 gene products, for which TAP-MS identified 415 interactors, 2HY screens identified 213 interactors, and only 28 proteins (13%) overlap between these two datasets. Similar percentage of overlapping interactors was reported in other independent studies [1, 170].

2HY screens detect pair-wise interactions of proteins, which is a significant limitation since many proteins are assembled into multi-subunit complexes stabilized by cooperative (rather than pair-wise) interactions. Therefore, many interactions

might be missed in two-hybrid screens if either bait or prey fails to assemble properly without other subunits of the complex.

Although TAP MS and 2HY screens provide a complementary view on physical interactions within the proteome, they do not correlate well enough to enable further validation of identified interactions (Figure 22). The concordance of results obtained by independent (although similar) protein tagging approaches [1, 2, 136], by 2HY methods [93, 94] or inferred *via* various bioinformatic approaches [167] is rather poor [171, 172]. Although the availability of complementary data is always a positive factor, it seems rather unlikely that observed discrepancies and excessive complexity of protein assemblies could be attributed solely to errors in analytical methods.

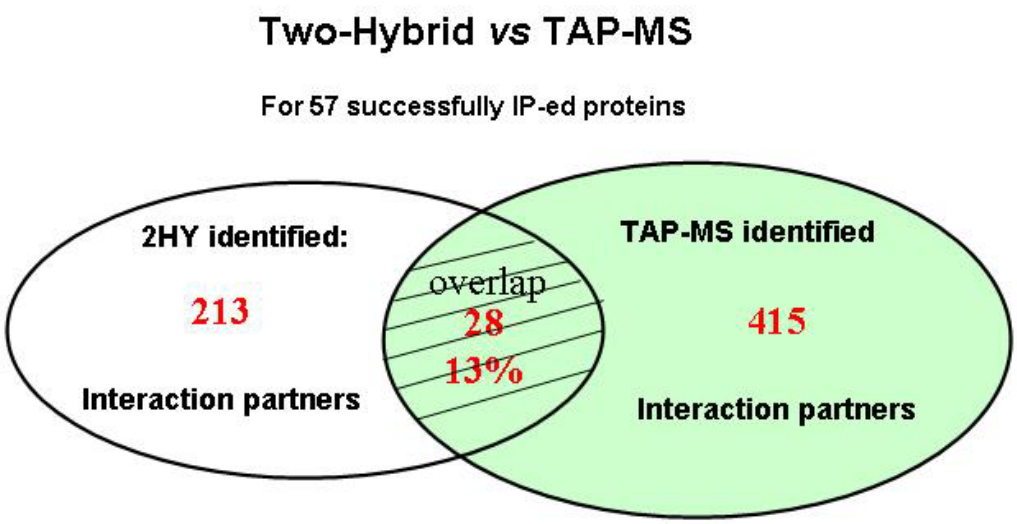


Figure 22. Overlap between interaction partners identified by TAP MS and by 2HY screens [136].

3.3. Dissection of protein complexes and proteomic hyperlinks in the budding and fission yeasts

Identification of interaction partners for a particular protein bait is only the first step in the characterization of the corresponding protein complex. This chapter elucidates how to correlate protein interactions discovered by TAP MS approach with biological experiments and properties of characterized protein assemblies and addresses the following specific questions:

- how to establish *bona fide* subunits of protein complexes by proteomics, rather than by labor-intense biochemical or genetics methods?
- how to learn if identified proteins belong to a single complex or to a mixture of several co-isolated protein complexes?
- how to chart a network of several protein complexes linked *via* shared subunits?
- how similar is the composition of protein complexes and their proteomic hyperlinks in phylogenetically distant organisms?
- how protein interactions discovered in premier model organisms *S. cerevisiae* and *S. pombe* could be projected on interactions in mammals, including humans?

3.3.1 Deciphering protein complexes *via* sequential protein tagging and mass spectrometry

We developed an experimental strategy based on **S**equential rounds of **E**pitope tagging, immuno**A**ffinity chromatography and **M**ass spectrometry (termed **SEAM** [170]) for dissecting the composition of individual protein complexes and charting segments of a protein interacting network [170, 173].

Typically, protein complexes are isolated *via* a single immunoaffinity purification step, in which one of already known subunits is used as a bait. After the identification of co-purified proteins, functional experiments should be performed to validate the discovered interactors and further define their function. These experiments (such as constructing deletion and temperature-sensitive mutants, co-localization with known proteins, etc) are very time consuming and often require expertise and skills that are beyond the precisely defined research focus of a molecular or cell biology laboratory [170, 174].

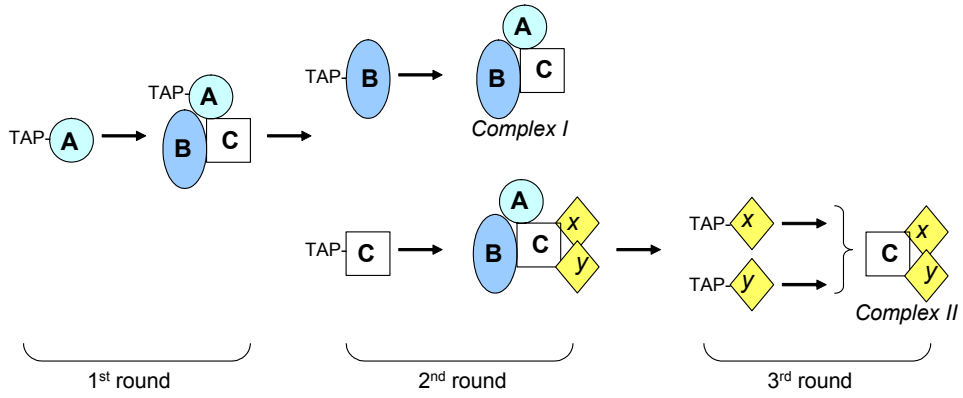


Figure 23. SEAM strategy. -In the first round a protein of interest (represented by A in green circle) is TAP-tagged and IP-ed, and its interaction partners are identified by mass spectrometry. However, this experiment does not provide direct evidence if the identified interaction partners belong to a single complex. To this end proteins identified in the 1st round are in turned tagged and IP-ed in 2nd round of SEAM. If these proteins were to form a stable complex, other subunits previously identified in the 1st round would be also identified (Subuits A, B and C, Complex I). However, some IPs might, together with already known subunits, produce other proteins, interacting only with given bait (yellow rombs *x* and *y*), but not with other members of the complex. Further IPs (3rd round) verify, if yet another individual complex has been identified (Complex II).

SEAM offers the advantage of systematic verification of identified interactions based on multiple and partially redundant purifications of the same protein assembly. In the first round of SEAM (Figure 23), a gene either encoding a known subunit of the complex, or selected because of bioinformatics or biological considerations, is tagged, and its interaction partners are sought. Interacting proteins identified in the first round are subsequently tagged and the procedure is repeated. It is conceivable that at some point subunit(s) associated with core subunits of other complexes would be identified thus linking the two complexes into a network. This approach is better suited for addressing specific biological problems *via* a comprehensive characterization of a given complex, rather than for global proteome-wide charting of the interactome.

By providing overlapping partially redundant sets of protein interactors, SEAM enables highly accurate charting of a proteomic environment *via* the

characterization of stable protein assemblies and the identification of subunits that are shared between two or more protein complexes, further termed "proteomic hyperlinks" [122, 123].

3.3.2 Dissection of Set3 complex in *S. cerevisiae*

SEAM strategy was applied to dissect the composition of the Set3C protein complex in the budding yeast and to characterize its hyperlinks [155]. Set3 (ORF name YKR029C) is one of the two proteins in *S. cerevisiae* whose sequences display both SET and PHD finger structural domains, which are hallmarks of proteins involved in chromatin regulation and epigenetics [155].

Previous publications implicated SET proteins [175, 176] and PHD finger proteins [177] in transcriptional silencing, however the latter proteins were also found among various transcriptional cofactors. No functional data were reported for Set3 protein in previous studies. *Δset3* strain lacked clear phenotype and there was no difference in viability compared with the wild type upon treatment with alkylating agents, UV irradiation, bleomycin, hydroxyurea, and heat [155]. To provide insight on Set3 function, TAP tag was fused to the C-terminus of the protein.

Immunoaffinity isolation of Set3 interacting proteins by the TAP method and subsequent mass spectrometric analysis identified eight specific proteins (Figure 24A). Among these eight proteins Kap95 and Kap60 belong to the family of importins and are reported to interact with each other [178]; Hos2p [179], and Hst1p [180] are putative histone deacetylases; the sequences of Sif2p [181] and YIL112w (later termed Hos4p) contain multiple repeats of generic protein-protein interaction motifs, WD40 and ankyrin; Snt1 protein contains putative DNA-binding SANT domain [182]; and Cph1p (cyclophilin A [183]) is a prolylisomerase.

The variability of cellular functions of Set3p interactors and the identification of two interacting histone deacetylases in a single protein assembly questioned the integrity and homogeneity of the isolated complex, since several distinct protein assemblies might have been inadvertently co-purified. To address this concern, Set3p interactors Sif2p, Snt1p, Hos4p, Cph1p, Hst1p, and Hos2p were tagged and subjected to the second round of TAP MS. Each time (with the exception of Cph1p, whose purification yielded only the bait) the same set of 7 proteins was co-IPed, thus

suggesting that they are *bona fide* members of a single complex termed Set3C (Figure 24B).

In further experiments, to dissect protein-protein interactions within Set3C, strains in which one complex member carried the TAP tag, whilst another subunit was deleted, were used (Figure 25). Attempts to purify intact Set3C from *Δset3* or *Δhos2* strains failed, suggesting the key role of these two proteins in the complex. At the same time, the intact Set3C was immunoprecipitated from *Δhst1* strain and only Hst1 protein was identified in Hos4-TAP IP in *Δhos2* strain. This indicated that the second histone deacetylase Hst1p directly interacts only with Hos4p within Set3C, whilst Hos2p is a *bona fide* core member of this complex.

Thus sequential rounds of TAP tagging and purification of Set3p and its interaction partners allowed us to identify and validate core subunits of the new protein complex termed Set3C. Furthermore, the identification of two putative histone deacetylases Hos2p and Hst1p among Set3p interactors navigated further experiments by providing a verifiable hypothesis on what the function of entire Set3C assembly might be. As was confirmed by further biological experiments, Set3C manifests NAD-dependent and NAD-independent deacetylase activities *in vitro* and also acts a repressor of meiosis [155].

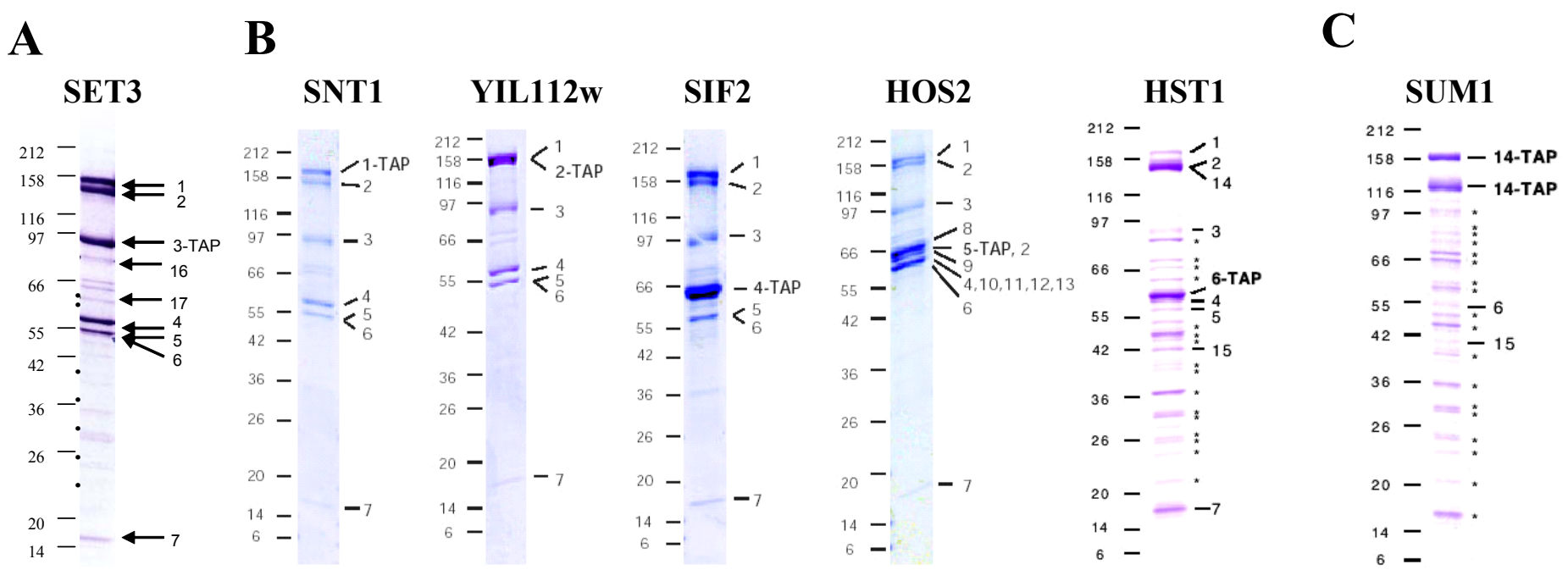


Figure 24. TAP MS dissection of Set3p proteomic environment. (A). TAP IP of the entry point Set3p, in which eight specific proteins were identified. (B). Second round of SEAM: tagging and immunoprecipitation of proteins identified in the first round. (C).. Third round of SEAM: Sum1-TAP immunoprecipitation confirmed Hst1p as a hyperlink to Sum1C. All visualized bands were analyzed by mass spectrometry. Bands of typical contaminants listed in Table 8A are designed with dots. Numbers designate: 1- Snt1p, 2- Hos4p, 3- Set3p, 4- Sif2p, 5- Hos2p, 6-Hst1p, 7- Cph1p, 8-13 – subunits of chaperonine complex TriC , 14- Sum1p, 15- Rfm1p, 16- Kap95p, 17- Kap60p [155].

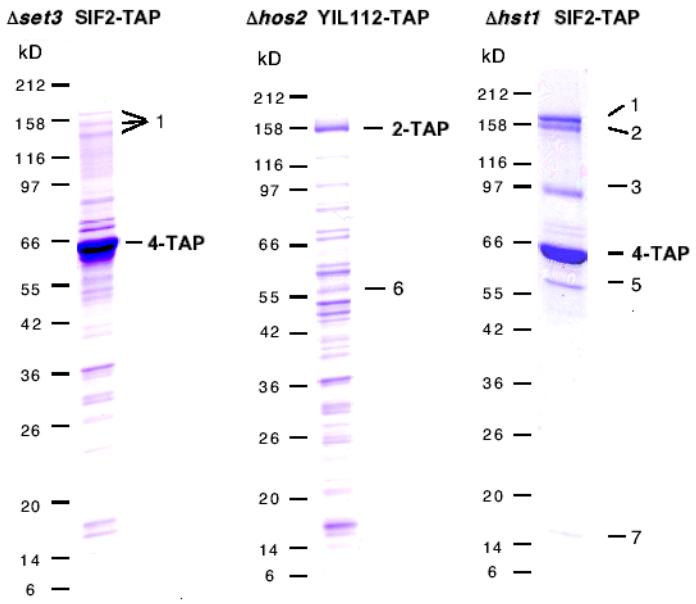


Figure 25. TAP isolation of Set3C from deletion strains. Same background proteins as presented in Table 9A were detected. Numbers at the right hand side indicate the same proteins as in Figure 24 [155].

3.3.3 TAP MS approach deciphers proteomic hyperlinks of Set3C complex.

As outlined in the section above, sequential tagging of interaction partners of Set3p confirmed that they form a stable protein complex. However, we also found that three subunits of Set3C were engaged in specific interactions with other protein assemblies. Importins Kap60p and Kap95p were pulled down with TAP-tagged Set3p only. Similarly, tagged Hos2p pulled down seven out of eight known subunits of the chaperonin complex TRiC [184] (which, unlike heat shock chaperones Ssa1p, Ssa2p, Ssb1p and Ssb2p, does not belong to common background proteins in TAP purifications). No importins and/or TriC chaperonins were pulled down by other TAP-tagged members of Set3C.

Tagging of Hst1p revealed that it is also engaged in another protein assembly with Rfm1p and Sum1p. To find out if these three proteins form yet another distinct previously uncharacterized complex, a third round of TAP tagging and purification was

performed on both Sum1p and Rfm1p (Figure 24C). We observed that both Sum1p and Rfm1p pulled down another two interactors (for example, Sum1p pulled down Rfm1p and Hst1p), but no other members of Set3C, and therefore we subsequently termed this complex Sum1C. Taken together with the results of immunoaffinity purifications from deletion strains (section 3.3.2, Figure 25), we identified Hst1p as a proteomic hyperlink between individual complexes Set3C and Sum1C (Figure 26).

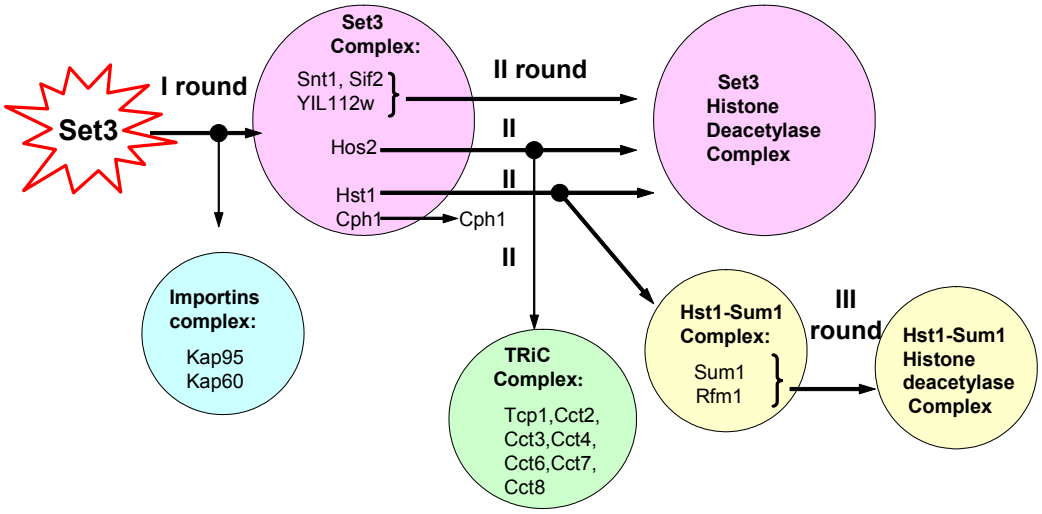


Figure 26. Set3p-anchored protein interaction network. Rounds of tagging are designated as I, II and III, individual complexes are circled. By starting at a single entry protein Set3p, SEAM approach enabled to identify two novel functionally distinct protein complexes with plausible histone deacetylase activity, linked via a shared subunit, Hst1p. SEAM rounds also revealed hyperlinks between two Set3C subunits (Hos2p and Set3p) and other known protein assemblies [136].

3.3.4 Set3p proteomic environment: comparison of sequential protein tagging *vs* genome-wide protein tagging

In this work TAP MS was initially applied to a small selection of interesting proteins and interaction partners identified in the first round were chosen for further rounds of tagging, as outlined in Figure 23 [136]. In an alternative experimental strategy, a large number of baits could be processed in parallel by an established high throughput protein purification and identification routine [1, 2]. The biological significance of identified interactions could be evaluated later and only for a selection of baits that yielded most interesting patterns of associated proteins. We further asked how accurately the genome-wide parallel analysis would have characterized the composition and hyperlinks of Set3C.

We used the results of the genome-wide parallel analysis available from Cellzome Yeast Database at http://yeast.cellzome.com to retrospectively assemble Set3p-proteomic environment (Figure 27) and to compare it with the environment dissected by SEAM (Figure 26). Genome-wide study also employed TAP MS as a method of the identification of interaction partners, although target genes were not selected sequentially, as in the SEAM strategy. Since only a fraction of the entire proteome was attempted and tagging of many of attempted genes failed, the obtained dataset was not at all comprehensive and therefore the composition of putative protein complexes was further outlined by bioinformatics. Three core subunits of Set3C, namely Sif2p, Cph1p and Hst1p, were among 1167 entry points presented in the Cellzome Yeast Database. In IPs with these baits CellZome team identified many proteins that were not detected in this study. All these proteins were merged into two complexes having different proposed functions: Complex 80, comprising 10 members, putatively involved in protein synthesis and turnover, and Complex 151, comprising 46 members, involved in signaling. Each putative complex was thought to comprise eight proteins, identified by us in Set3p proteomic environment. Sif2p and Hst1p were also included in yet another Complex 188 that differed from Complex 80 and Complex 151.

Seven proteins from Set3p proteomic environment that belonged to TRiC complex and importins were almost randomly detected in various IPs and were assigned

to 18 different complexes, comprising from three to 43 members. For example, Kap95p and Kap60p were assigned to Complex 129, which comprised 35 members, and was supposed to be involved in RNA methabolism. Although TRiC complex has been extensively characterized in the budding yeast as a complex of eight individual proteins [184], not a single complex from 18 reported in the Cellzome database comprised all of its known subunits. Members of Sum1C were not identified in any IP.

We therefore concluded that the characterization of Set3p proteomic environment provided by both genome-wide projects lacked the accuracy and did not corroborate with evidences acquired by independent methods.

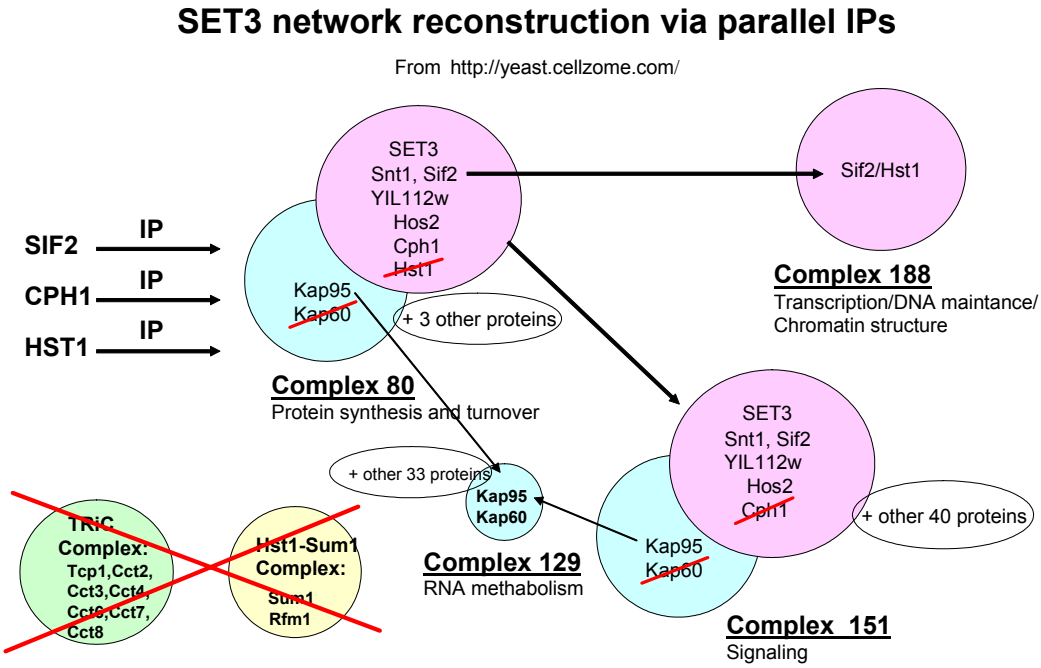


Figure 27. Set3p anchored protein network reconstruction via partial parallel tagging. Three subunits Sif2p, Cph1p and Hst1p were reported to be TAP-tagged and IPed. The source of raw data: www.yeast.cellzome.com (see section 3.3.4 and Figure 26 for details).

3.3.5 Proteomic environments in *S. cerevisiae* determined by various methods

We compared in the same way several proteomics environments deciphered by us in *S. cerevisiae*. To this end, the results of TAP IPs of 18 baits from the list presented in Table 9 (Set2p, Hst1p, Sir2p, Sif2p, Cph1p, Esa1p, Swd1p, Swd3p, Reb1p, Chd1p, Epl1p, Isw1p, Rpd3p, Sin3p, Hda1p, Zds1p, Nam8p and Rse1p) were compared with corresponding raw data from Cellzome Yeast Database (Figure 28).

For the same dataset of 18 genes the CellZome genome-wide screen revealed 162 interaction partners, whilst 117 interactors were identified *via* SEAM apparoach. 63 proteins (39%) overlap between the two dataset.

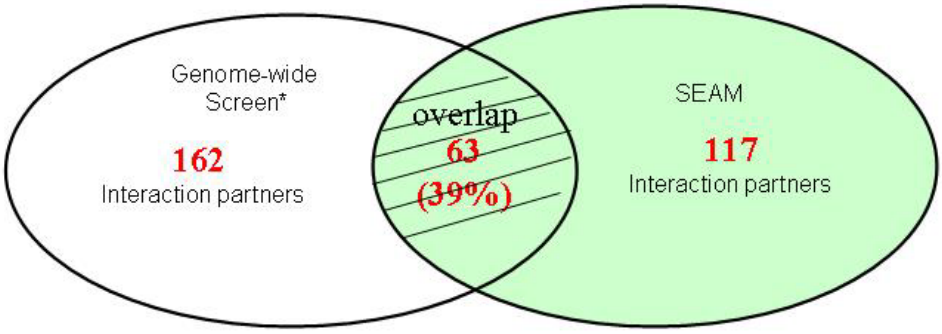


Figure 28. Comparison of interaction partners identified in genome-wide screen and by SEAM approach. * from CellZome Yeast Database http://yeast.cellzome.com

99 interaction partners identified in the genome-wide study were not observed in SEAM rounds. At the same time, genome-wide database missed 54 true interaction partners identified and validated by SEAM and by subsequent functional experiments.

The major difference in the methodology between high throughput charting of protein-protein interactions [1, 2] and SEAM approach is that the latter produced a highly redundant dataset, in which newly identified interactors were detected several times in independent purification experiments. Also the sequential tagging strategy allowed us to pinpoint proteins shared by distinct complexes and therefore reduced the risk of erroneousmerges of otherwise individual protein assemblies. The sequential strategy reduced to absolute minimum the number of false-positive interactors and this might

explain why SEAM could not confirm more than 2/3 of interactors identified in high-throughput studies.

We note here that advanced bioinformatics methods were not able to amend the lack of protein-protein interaction data. We demonstrated in the section above that it was not possible to accurately recreate Set3p proteomic environment from interactions discovered by genome-wide screens. Parallel tagging strategy failed to distinguish core members of Set3C complex that are essential for its integrity, from proteomic hyperlinks. We next asked if such poor concordance also occurs in the comparative analysis of protein complexes in the budding yeast.

To answer these questions, a large network of proteins involved in chromatin regulation, which consists of more than 50 members assembled in five hyperlinked complexes was characterized by SEAM. Altogether, 17 baits were TAP-tagged in sequential manner followed by the identification of their interaction partners by a combination of mass spectrometric techniques. The determined proteomic environments were compared with protein complexes and hyperlinks suggested by the two genome-wide studies [1, 2]. Figure 29 demonstrates that protein complexes reported by Gavin *et al.* [1] and Ho *et al.* [2] are in poor concordance and proteomic hyperlinks connecting individual protein assemblies are completely missed.

The poor concordance might be only partially attributed to apparent technical limitations in strain preparation, protein purification or mass spectrometric identification. It is more important that experimental strategy tuned to high proteome coverage and high throughput was not able to produce data of sufficient redundancy, which could have been used as internal consistency controls. Relatively low (70 %) probability of detecting the same protein in two independent purifications of the same entry protein [1] was noted. It means that, on average, 30% of all associations detected in high-throughput studies should be treated with caution and verified by some independent methods. This is particularly important if a protein complex was characterized *via* IP of only one entry protein.

Since many IPs pool together members of different protein complexes, “guilty by association” concept of defining what proteins belong to the complex looks inherently

error-prone. This is a yet another severe deficiency of parallel tagging strategies, in which *bona fide* interaction partners are established neither by functional experiments nor by sequential tagging and purification of other candidate subunits of the complex. Our data indicate that almost a quarter of budding yeast proteins is shared between two or more protein complexes and therefore multiple independent purifications of each complex using different baits are necessary.

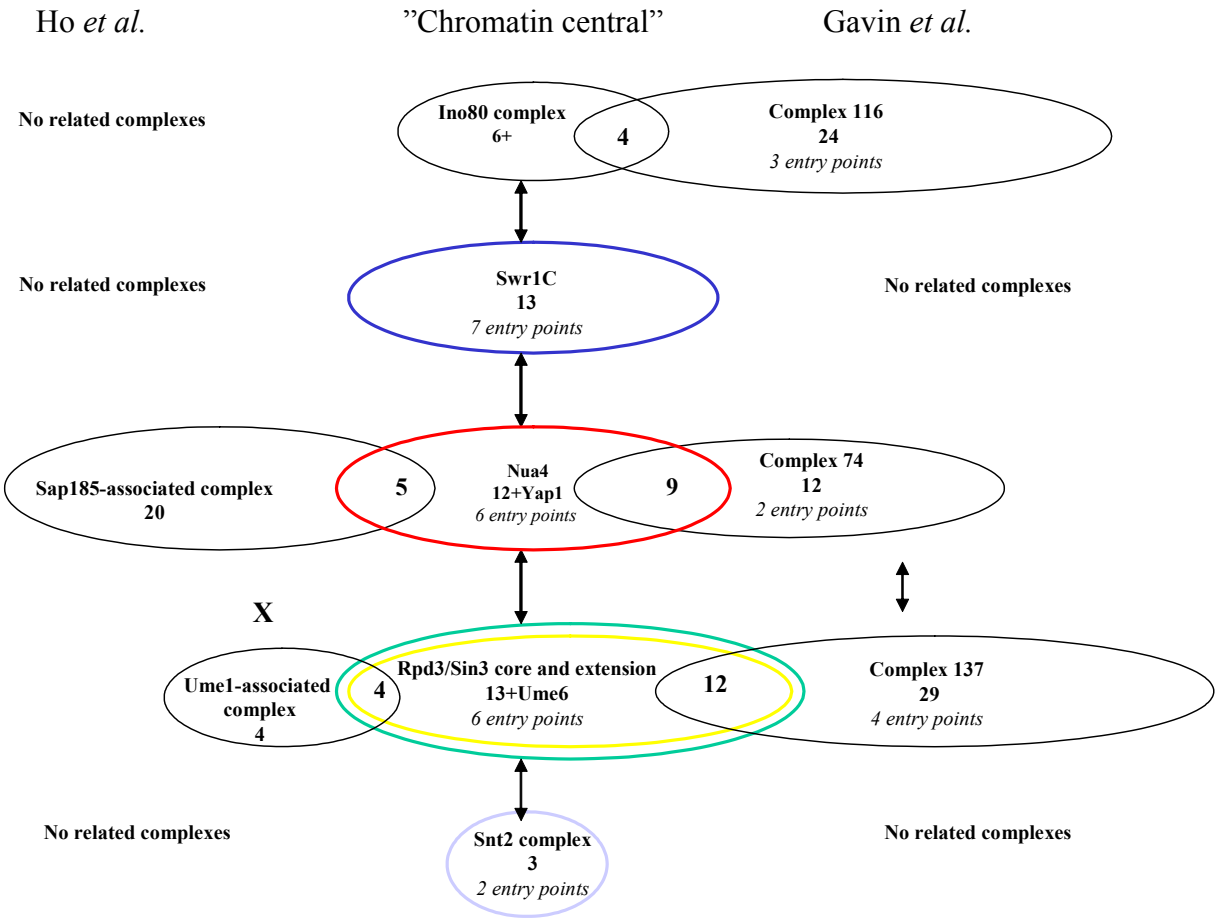


Figure 29. Large protein interaction network in *S. cerevisiae* by different approaches. From the left hand side: from Ho *et al*. [2]; in the middle by TAP-MS and SEAM (this study); from the right-hand side: from Gavin *et al*. [1].

3.3.6 Comparative proteomic analysis of orthologous proteomic environments

Since the sequences and function of many proteins are conserved among a variety of species, it has been inferred that orthologous protein complexes might also share similar composition and architecture. Thus it is conceivable that conserved protein complexes could be initially characterized in a model organism and then the obtained knowledge might be projected on orthologous complexes in other organisms, including humans. However it is not clear, if the composition of orthologous non-essential protein complexes (e.g. complexes, which could be disrupted without causing a lethal phenotype) is also conserved, and if the conservation also extends to proteomics hyperlinks. To address these issues, we dissected and compared Set1p orthologous proteomic environments in the budding and fission yeast by TAP-MS approach.

3.3.6.1 Set1 proteomic environment in *S. cerevisiae*

Set1 complex is termed after its key subunit Set1p (ORF name YHR119w), whose sequence possesses a characteristic SET-domain [176]. The complex methylates lysine 4 in histone H3 and is implicated in epigenetic regulation [137, 185].

A protein complex comprising Set1 protein was identified by its C-terminal TAP tagging and the identification of co-purified proteins by MALDI peptide mass fingerprinting. All Coomassie stained bands were identified, however, only seven proteins (Spp1p; Bre2p; three WD40 repeat proteins Swd1p, Swd2p, Swd3p; Shg1p and Sdc1p) were recognized as *bona fide* members of Set1C and the rest were common background proteins (see Table 9A).

These seven proteins were subsequently TAP tagged and IPed and six of them pulled down other seven proteins with similar stoichiometry, as was judged by relative intensity of Coomassie staining of corresponding protein bands (Figure 30A). However, the seventh protein, Swd2p, was the notable exception. The pool of proteins co-isolated with Swd2p included all members of Sc_Set1C and 9 members of another yeast complex termed CPF for Cleavage and Polyadenylation Factor (Figure 31A, Table 14).

Two other groups independently confirmed that Swd2p is a true member of the budding yeast CPF, however its association with Set1C was not reported [1, 186]. Using two entry points, Swd1 and Swd3 proteins, Gavin *et al.* [1] identified a complex (termed Complex 108) having the protein composition similar to Set1C. The Complex 108 missed two subunits (Swd2p and Shg1p) and, consequently, a hyperlink to CPF complex *via* Sc_Swd2p. At the same time, Complex 108 comprised three other proteins, whose relation to Sc_Set1C was not independently confirmed. Gavin *et al.* also tagged seven out of 20 known subunits of CPF and detected Swd2p in all affinity purifications [1]. However, Sc_Swd2p itself was not tagged and its relation to Sc_Set1C was not established.

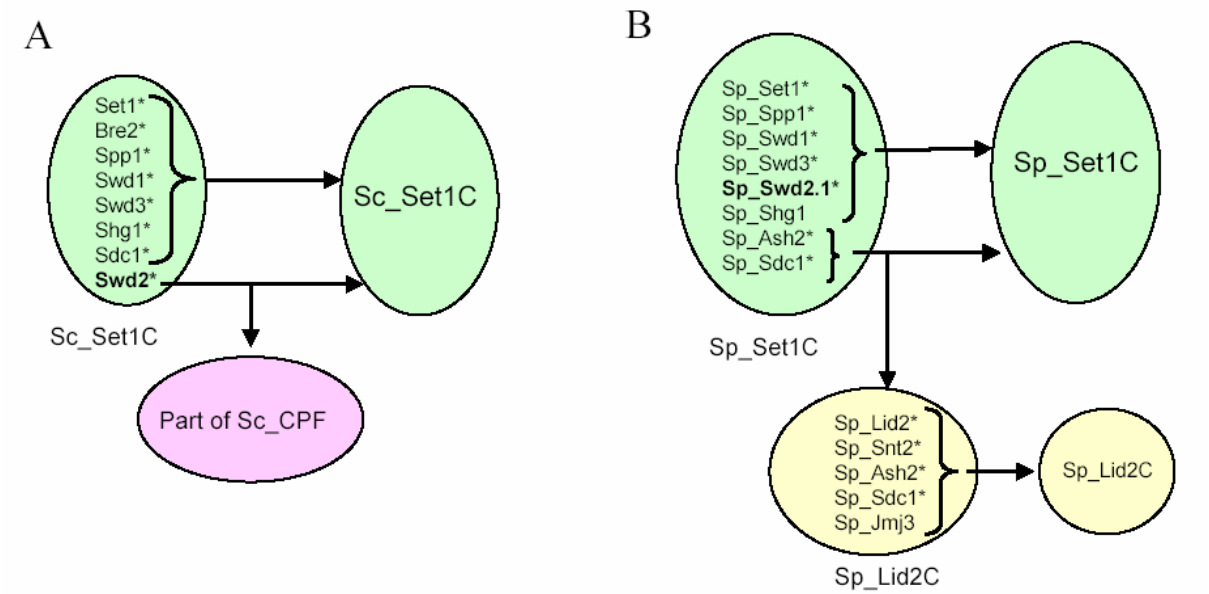


Figure 30. Proteomic environment of Set1 proteins in *S. cerevisiae* (panel A) and *S. pombe* (panel B). Individual complexes are circled. Successfully TAP purified proteins are indicated with asterisks. Sc_Set1C in comprises 8 subunits, including Sc_Swd2p identified as a hyperlink to Sc_CPF. Sp_Set1C also comprises 8 subunits. Sp_Ash2p and Sp_Sdc1p are hyperlinks between Sp_Set1C and a new protein complex termed Sp_Lid2C [123].

Taken together, these data suggested that Swd2 protein is a subunit shared between two independent complexes, Sc_Set1C and Sc_CPF and that it hyperlinks histone methylation and polyadenylation machinery in the budding yeast. Although both complexes act at the site of active transcription, the significance of this hyperlink still remains elusive. At the same time, sequential rounds of epitope tagging confirmed eight proteins, Set1p, Spp1p, Bre2p, Swd1p, Swd2p, Swd3p, Shg1p and Sdc1p to be *bona fide* members of Sc_Set1C.

3.3.6.2 Dissection of Set1p proteomic environment in *S. pombe* and its comparison with the orthologous proteomic environment in *S. cerevisiae*

As a next step, TAP MS was applied for the characterization of the Set1p proteomic environment in *S. pombe* and the composition of orthologous complexes Sc_Set1C and Sp_Set1C and their hyperlinks were compared.

The genome of *S. pombe* encodes for the orthologous protein Sp_Set1p (ORF name SPCC306.04c), which shares 29 % of the full-length sequence identity with its *S. cerevisiae* homologue. This protein was selected as an entry to purify Sp_Set1C. We found that, similarly to Sc_Set1C, Sp_Set1C also comprises 8 subunits, which (with the exception of Sp_Shg1p) were the closest homologues of the corresponding members of Sc_Set1C [122, 123]. Notably, the fact that these complexes are conserved between *S. cerevisiae* and *S. pombe* could not be deduced by bioinformatic extrapolation from the known composition of Sc_Set1C. In particular, TAP MS identified Sp_Set1C subunits Sp_Shg1p, Sp_Ash2p, Sp_Swd2.1p and Sp_Swd3p that lack convincing sequence homology to corresponding members of Sc_Set1C (Table 14).

The characterization of Sp_Set1C did not detect its interaction with Sp_CPF. By immunoprecipitating tagged members of Sp_Set1C, we established that Sp_Swd2.1p (ORF name SPBC18H10.06c) is a member of Sp_Set1C (Figure 31B, Table 14). However, IP of Sp_Swd2.1p did not provide any evidence that it interacts with Sp_CPF complex. The genome of *S. pombe* encodes for yet another distant homologue of Sc_Swd2p, namely SPAC824.04 (now termed Sp_Swd2.2p), sharing 32% of sequence identity with Sc_Swd2p and 30 % of sequence identity to Sp_Swd2.1p.

We therefore investigated if another paralogue of Sc_Swd2p, Sp_Swd2.2p is a yet undetected hyperlink between Sp_Set1C and Sp_CPF. Sp_Swd2.2p was TAP tagged and its IP pulled down the Sp_CPF complex but, importantly, no members of Sp_Set1C complex were detected (Figure 31B). We further attempted to purify Sc_CPF *via* tagging and IP of one of its conserved subunits, Sp_Ysh1p (ORF name SPAC17g6.16c) and checked if one or both Sc_Swd2p paralogues would be co-isolated with known core subunits of the complex. In this IP experiment we identified the stoichiometric quantity of Sp_Swd2.2p compared to another 11 members of Sp_CPF, but no Sp_Swd2.1p was purified in a detectable amount. Taken together, these data suggested that Sp_Swd2.2p, but not Sp_Swd2.1p is a genuine member of Sp_CPF. We conclude that duplicated *SWD2* genes in *S. pombe* are functionally specialized [187, 188] with Sp_Swd2.1 protein being a member of Sp_Set1C and Sp_Swd2.2 being a member of Sp_CPF (Figure 31).

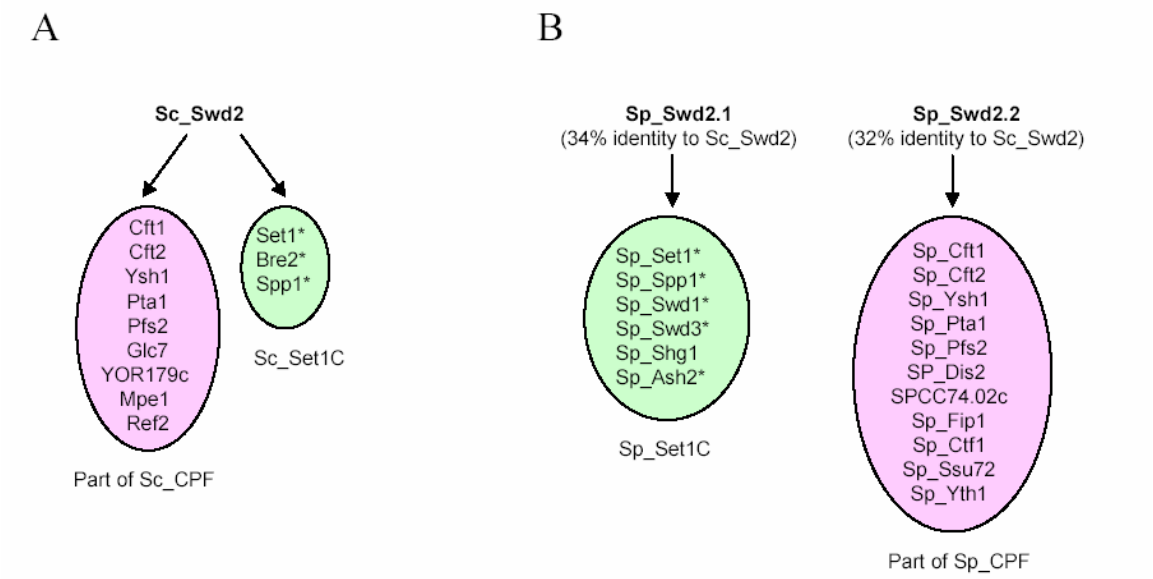


Figure 31. Interaction partners of Swd2p in *S. cerevisiae* (Panel A) and of its paralogues in *S. pombe* (Panel B). In *S. cerevisiae*, Sc_Swd2p hyperlinks Sc_SetC1 and Sc_CPF complexes. In *S. pombe* the function of the Swd2p homologues is fully diverged, with Sp_Swd2.1p becoming a member of Sp_Set1C and Sp_Swd2.2p becoming a member of Sp_CPF [123].

Thus comparative study of the accurately charted Set1p proteomic environments established that a hyperlink between Set1C and CPF is present in *S. cerevisiae*, but is absent in *S. pombe*.

At the same time, the characterization of Sp_Set1C revealed a novel proteomic hyperlink, which was not observed in *S. cerevisiae* (Figure 30B). Along with members of Sp_Set1C, both Sp_Ash2p and Sp_Sdc1p pulled down yet another set of interacting proteins consisting of Sp_Jmj3p and two PHD-finger proteins Sp_Snt2p and Sp_Lid2p (Table 14). None of these three proteins were co-isolated when other Sp_Set1C subunits were TAP tagged. Another round of protein tagging and IP was performed using Sp_Lid2p and Sp_Snt2p as baits and established the presence of a new complex termed Sp_Lid2C (Figure 30B).

No protein complex that is orthologous to Sp_Lid2C was found in the budding yeast, despite the presence of a few reasonable sequence homologues in its genome. The genome of *S. cerevisiae* encodes for proteins YJR119c and Snt2p (YGL131c), which share

Table 14.
Composition of protein complexes from *S. cerevisiae* and *S. pombe*

S.cerevisiae[a]			***S.pombe***			
Gene	ORF	MW, kDa	Gene	ORF	MW, kDa	Identity / Similarity, %
	Sc_Set1C			**Sp_Set1C**		
Set1	YHR119w	124	Set1	SPCC306.04c	105	29 / 44
Bre2	YLR015w	58	Ash2	SPBC13g1.08c	74	27 / 39
Spp1	YPL138c	41	Spp1	SPCC594.05c	49	40 / 58
Swd1	YAR003w	49	Swd1	SPAC23h3.05c	45	37 / 55
Swd2	YKL018w	37	Swd2.1	SPBC18h10.06c	40	34 / 53
Swd3	YBR175w	35	Swd3	SPBC354.03	43	30 / 49
Shg1	YBR258c	16	Shg1	SPAC17g8.09	15	-
Sdc1	YDR469w	19	Sdc1	SPCC18.11c	12	46 / 75
	Part of Sc_CPF			**Part of Sp_CPF**		
Cft1	YDR301w	153	Cft1	SPBC1709.08	160	25 / 45
Cft2	YLR115w	96	Cft2	SPBC1709.15c	89	25 / 44
Ysh1	YLR227c	88	Ysh1	SPAC17G6.16c	88	49 / 67
Pta1	YAL043c	88	Pta1	SPAC4H3.15c	50	23 / 40
Swd2	YKL018w	37	Swd2.2	SPAC824.04	38	32 / 55
Pfs2	YNL317w	53	Pfs2	SPAC12G12.14c	58	43 / 61
Glc7	YER133w	36	Dis2	SPBC776.02c	37	89 / 95
			Yth1	SPAC227.08c	19	
			Ssu72	SPAC3G9.04	23	
			Fip1	SPAC22G7.10	37	
			Cft1	SPBC3b9.11c	40	
Ref2	YDR195w	60				-
Mpe1	YKL059c	50				
	YOR179c	21	Ysh1	SPAC17G6.16c	88	26 / 48
				SPCC74.02c	77	-
				Sp_Lid2C		
			Lid2	SPBP19A11.06	172	
			Snt2	SPAC3h1.12c	129	22 / 38
			Jmj3	SPBC83.07	85	

23% and 22% of full-length sequence identity with Sp_Lid2p and Sp_Snt2p, respectively and display a very similar composition of functional domains. Sc_Snt2p and YJR119C were TAP tagged and their interaction partners were successfully isolated, but no proteins even distantly homologous to members of Sp_Lid2C were identified [123].

Thus the comparative proteomic analysis prompted us to conclude that Lid2C and its hyperlink to Set1C are specific for *S. pombe* and are not present in *S. cerevisiae*. Although the composition of Set1C is conserved between *S. cerevisiae* and *S. pombe*, the hyperlinks of Set1C complex are different in two ways: first, in the fission yeast Sp_Set1C is hyperlinked to another complex Sp_Lid2C, which does not exist in the budding yeast. Second, Set1C and CPF are no longer hyperlinked *via* Swd2 - homologous proteins in *S. pombe*.

3.3.6.3 Human Set1C: bioinformatic predictions and experimental evidences

Using TAP MS method, we characterized in detail the Set1p proteomic environments in the budding and fission yeasts. Based on this knowledge and on recently published evidences, we next tried to project the molecular organization of Set1p orthologous environment in humans.

The human genome encodes for two putative proteins homologous to Set1p in both fungi. The proteins KIAA0339 and KIAA1076 share 35% and 37% of sequence identity to Sc_Set1p and 55% and 45% identity to Sp_Set1p. The KIAA0339 is engaged in a partially characterized complex comprising at least hAsh2p and Wdr5p (a human homologue of Sc_Swd3p) [189] and is also involved in H3K4 methylation. These proteins were found by IP with antibodies raised against human protein HCF-1, which has no apparent homology to any of the core members of Sp_Set1C or Sc_Set1C, but is known to associate with human Sin3 histone deacetylase (HDAC). We could therefore speculate that the partial purification of human Set1C –related protein assembly was probably achieved *via* a hyperlink protein.

Comparing *S. cerevisiae* Set1C with partially purified putative human Set1C, Wysocka *et al.* [189] proposed that it might contain no orthologues of Sc_Spp1p – a protein with a PHD finger domain (although human genome encodes for CGBP, a

reasonably close homologue of Sc_Spp1p [137]), but its absence is functionally compensated by a PHD finger domain in the sequence of hAsh2p (Figure 32). Similar to the orthologous complex Sc_Set1C, the fission yeast Sp_Set1C is also composed of eight proteins, however the sequences of two of them - Sp_Spp1p and Sp_Ash2p - comprise a PHD finger domain. Thus we hypothesized that the human homologue of Spp1p could be a *bona fide* member of the orthologous Set1C complex in humans. This hypothesis could be validated either by direct purification of the complex *via* the human homologue of Spp1p (CGBP protein) or by multiple purifications *via* known subunits of human Set1C.

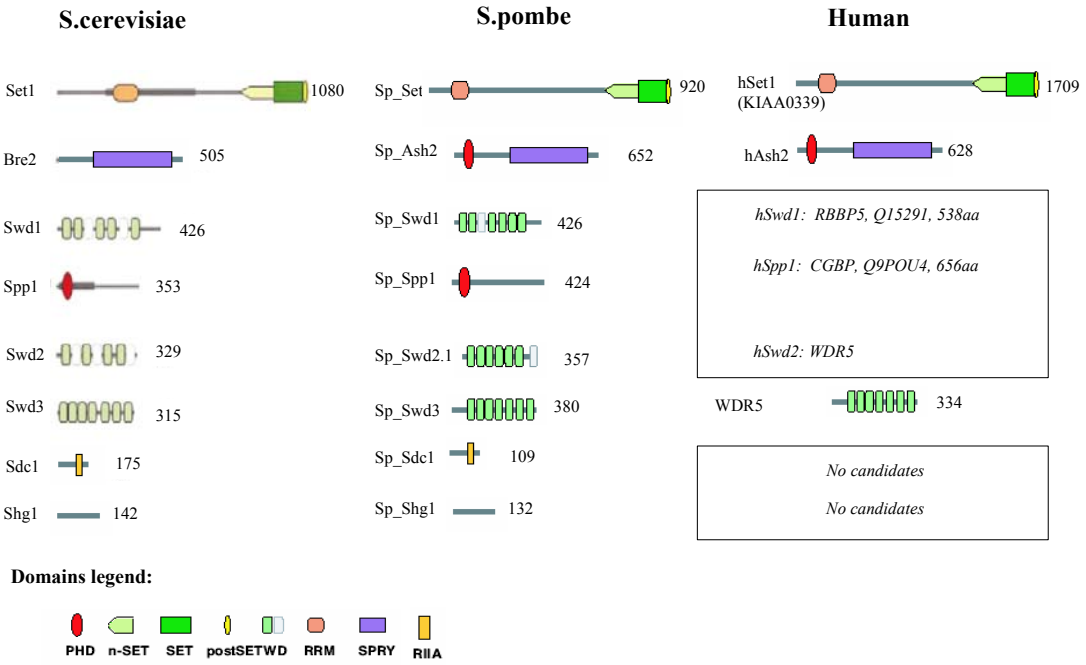


Figure 32. Comparison of the domain composition of Set1C subunits in different species. Domain composition of *S. cerevisae* Set1C is from Roguev *et al.* [137], *S. pombe* Set1C is from Roguev *et al.* [122], putative human Set1C subunits are from Wysocka *et al.* [189]. Predictions are in *italic* in squares (results of BLAST searches were taken from HumanPD, www.incyte.com).

Taken together, the composition of Sc_Set1C and Sp_Set1C and experimental evidences from partial purification of the orthologous complex in humans indicate that although the overall composition of human Set1C may be similar, its proteomic environment comprises (at least) one novel hyperlink, HCF-1 protein.

3.3.6.4. Protein complexes in the phylogenetic perspective

Our results underscore the value and importance of accurate characterization of the composition of protein complexes *via* multiple independent purifications using different baits, especially when their proteomic environments are considered in the phylogenetic perspective. The sequential approach applied in this work was rather laborious: altogether, for 3 protein complexes characterized in *S. pombe* (Sp_Set1C, Sp_Lid2C and Sp_CPF) 12 baits were TAP- tagged and mass spectrometric identification of 681 bands of co-isolated proteins was performed (Table 13), so that each complex was independently purified several times using different subunits as baits. However, SEAM approach minimizes the risk of missing a core subunit of the complex or erroneously merging of otherwise distinct protein complexes and these are key features for understanding of the molecular architecture of proteomic environments.

Taken together, our data and other published evidences [144] strongly suggest that, although orthologous protein complexes may be remarkably conserved, their proteomic environment and hyperlinks to other complexes are not. Furthermore, the conservation of the core and variability of links may represent a common phenomenon in the molecular organization of eukaryotic proteomes. The comparative analysis of proteomic environments in a multiorganismal perspective offers an intriguing opportunity to extend and complement our understanding of how the evolution of genomes guides the evolution of protein machines. Comparative studies may reach far beyond simple cataloguing of observed differences. Rather, together with advanced bioinformatic approaches, correlations of concerted alterations in sequences of orthologous subunits could highlight functional specializations.

4. Summary

Characterization of protein complexes by mass spectrometry based proteomics presents numerous analytical challenges. Key features of the employed analytical routine, including the purification of complexes and mass spectrometric identification of their subunits were investigated in details. Archiving, storage and handling of polyacrylamide gels, visualization of protein bands and their effect on the efficiency of *in-gel* digestion and mass spectrometric identification of proteins were quantitatively evaluated. It was further demonstrated that a combination of several mass spectrometric techniques based on MALDI and ES ionization provides highly complementary data and enable comprehensive characterization of protein digests.

The accumulated knowledge and optimized analytical procedures assisted in developing the experimental approach for deciphering protein complexes and protein interaction networks in the budding and fission yeasts. The results presented in this work demonstrated that a combination of Tandem Affinity Purification (TAP) and mass spectrometric identification of gel-separated protein subunits is a generic and robuststrategy that provides accurate and reproducible data.

Analytical aspects of the TAP MS strategy were elucidated. The evaluation of its success rate, reproducibility and typical protein background presented in this work is based on TAP tagging and immunoprecepitation of 75 genes in *S. cerevisiae* and 22 in *S. pombe*. The molecular composition of the characterized protein complexes was compared with the protein-protein interactions uncovered by other established methods, such as yeast two hybrid screens or proteome-wide purification of protein complexes. It has been demonstrated that repetitive purification of protein complexes using different subunits as baits is crucially important for confident charting of proteomics environments. Accurate dissection of the composition of individual protein complexes and identification of their proteomic hyperlinks enabled to consider proteomic envitionments in the phylogentic perspective and paved the way to reliable projection of proteomics data obtained in lower eukaryotic model organisms to higher eukaryotes, including humans.

5. REFERENCES

1 Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.

2 Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183.

3 Neubauer, G., Gottschalk, A., Fabrizio, P., Seraphin, B., Luhrmann, R., Mann, M. (1997) Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc Natl Acad Sci U S A* **94**, 385-390.

4 Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P. *et al.* (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet* **20**, 46-50.

5 Rout, M.P., Aitchison, J.D., Suprapto, A., Hjertaas, K., Zhao, Y., Chait, B.T. (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J Cell Biol* **148**, 635-651.

6 Zhou, Z., Licklider, L.J., Gygi, S.P., Reed, R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182-185.

7 Taylor, S.W., Fahy, E., Ghosh, S.S. (2003) Global organellar proteomics. *Trends Biotechnol* **21**, 82-88.

8 Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570-574.

9 Wigge, P.A., Jensen, O.N., Holmes, S., Soues, S., Mann, M., Kilmartin, J.V. (1998) Analysis of the *Saccharomyces* spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. *J Cell Biol* **141**, 967-977.

10 Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W. *et al.* (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537-542.

11 Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D. *et al.* (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520-526.

12 Foster, L.J., De Hoog, C.L., Mann, M. (2003) Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc Natl Acad Sci U S A* **100**, 5813-5818.

13 von Haller, P.D., Donohoe, S., Goodlett, D.R., Aebersold, R., Watts, J.D. (2001) Mass spectrometric characterization of proteins extracted from Jurkat T cell detergent-resistant membrane domains. *Proteomics* **1**, 1010-1021.

14 Karas, M., Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301.

15 Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71.

16 Beavis, R.C., Chaudhary, T., Chait, B.T. (1992) α-Cyano-4-hydroxycinnamic acid as a matrix for matrix assisted laser desorption mass spectrometry. *Org. Mass Spectrom.* **27**, 156-158.

17 Horneffer, V., Forsmann, A., Strupat, K., Hillenkamp, F., Kubitscheck, U. (2001) Localization of analyte molecules in MALDI preparations by confocal laser scanning microscopy. *Anal Chem* **73**, 1016-1022.

18 Gluckmann, M., Pfenninger, A., Kruger, R., Thierolf, M., Karas, M., Horneffer, V. *et al.* (2001) Mechanisms in MALDI analysis: surface interaction or incorporation of analytes? *Int J Mass Spectrom* **210**, 121-132.

19 Aebersold, R., Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198-207.

20 Yates, J.R., 3rd (2000) Mass spectrometry. From genomics to proteomics. *Trends Genet* **16**, 5-8.

21 Xu, Y., Bruening, M.L., Watson, J.T. (2003) Non-specific, on-probe cleanup methods for MALDI-MS samples. *Mass Spectrom Rev* **22**, 429-440.

22 Vorm, O., Roepstorff, P., Mann, M. (1994) Matrix surfaces made by fast evaporation yield improved resolution and very high sensitivity in MALDI TOF. *Anal. Chem.* **66**, 3281-3287.

23 Shevchenko, A., Wilm, M., Vorm, O., Mann, M. (1996) Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850-858.

24 Jensen, O.N., Mortensen, P., Vorm, O., Mann, M. (1997) Automatic acquisition of MALDI spectra using fuzzy logic control. *Anal. Chem.* **69**, 1706 - 1714.

25 Yost, R.A., Boyd, R.K. (1990) Tandem mass spectrometry: quadrupole and hybrid instruments. *Methods Enzymol* **193**, 154-200.

26 Shevchenko, A., Chernushevich, I., Wilm, M., Mann, M. (2000) De Novo peptide sequencing by nanoelectrospray tandem mass spectrometry using triple quadrupole and quadrupole - time -of-flight instruments. *Meth Mol Biol* **146**, 1-16.

27 Wilm, M., Neubauer, G., Mann, M. (1996) Parent ion scans of unseparated peptide mixtures. *Anal Chem* **68**, 527-533.

28 Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T. *et al.* (1996) Femtomole sequencing of proteins from polyacrylamide gels by nanoelectrospray mass spectrometry. *Nature* **379**, 466-469.

29 Wilm, M., Mann, M. (1996) Analytical properties of the nano electrospray ion source. *Anal. Chem.* **66**, 1-8.

30 Mann, M., Wilm, M. (1995) Electrospray mass spectrometry for protein characterization. *Trends Biochem Sci* **20**, 219-224.

31 Chernushevich, I., Loboda, A., Thomson, B. (2001) An introduction to quadrupole time-of-flight mass spectrometry. *J. Mass Spectrom.* **36**, 849-865.

32 Chernushevich, I. (2000) Duty cycle improvement for a quadrupole time-of-flight mass spectrometer and its use for precursor ion scans. *Eur J Mass Spectrom* **6**, 471-479.

33 Krutchinsky, A.N., Loboda, A.V., Spicer, V.L., Dworschak, R., Ens, W., Standing, K.G. (1998) Orthogonal injection of matrix-assisted laser

desorption/ionization ions into a time-of-flight spectrometer through a collisional damping interface. *Rapid Commun Mass Sp* **12**, 508-518.

34 Krutchinsky, A.N., Zhang, W., Chait, B.T. (2000) Rapidly switchable matrix-assisted laser desorption/ionization and electrospray quadrupole-time-of-flight mass spectrometry for protein identification. *J Am Soc Mass Spectrom* **11**, 493-504.

35 Loboda, A.V., Krutchinsky, A.N., Bromirski, M., Ens, W., Standing, K.G. (2000) A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption / ionization source: design and performance. *Rapid Commun. Mass Spectrom.* **14**, 1047-1057.

36 Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., Standing, K.G. (2000) MALDI Quadrupole Time-of-Flight mass spectrometry: a powerful tool for proteomic research. *Anal. Chem.* **72**, 2132-2141.

37 March, R.E. (1997) An introduction to quadrupole ion trap mass spectrometry. *J Mass Spectrom* **32**, 351-369.

38 Krutchinsky, A.N., Kalkum, M., Chait, B.T. (2001) Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Anal Chem* **73**, 5066-5077.

39 Hager, J.W. (2002) A new linear ion trap mass spectrometer. *Rapid Commun Mass Sp* **16**, 512-526.

40 Schwartz, J.C., Senko, M.W., Syka, J.E. (2002) A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* **13**, 659-669.

41 Martin, S.E., Shabanowitz, J., Hunt, D.F., Marto, J.A. (2000) Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* **72**, 4266-4274.

42 Medzihradszky, K.F., Campbell, J.M., Baldwin, M.A., Falick, A.M., Juhasz, P., Vestal, M.L. *et al.* (2000) The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal Chem* **72**, 552-558.

43 Horn, D.M., Zubarev, R.A., McLafferty, F.W. (2000) Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc Natl Acad Sci U S A* **97**, 10313-10317.

44 Gomez, S.M., Nishio, J.N., Faull, K.F., Whitelegge, J.P. (2002) The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. *Mol Cell Proteomics* **1**, 46-59.

45 Rabilloud, T. (2000) Detecting proteins separated by 2-D gel electrophoresis. *Anal Chem* **72**, 48A-55A.

46 Washburn, M.P., Wolters, D., Yates, J.R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-247.

47 Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**, 676-682.

48 Schirle, M., Heurtier, M.A., Kuster, B. (2003) Profiling Core Proteomes of Human Cell Lines by One-dimensional PAGE and Liquid Chromatography-Tandem Mass Spectrometry. *Mol Cell Proteomics* **2**, 1297-1305.

49 Fenyo, D. (2000) Identifying the proteome: software tools. *Curr Opin Biotechnol* **11**, 391-395.

50 Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A* **90**, 5011-5015.

51 Pappin, D.J.C. (2002) Peptide mass fingerprinting using MALDI-TOF mass spectrometry. *Meth Mol Biol* **211**, 211-217.

52 Jensen, O.N., Podtelejnikov, A.V., Mann, M. (1997) Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* **69**, 4741-4750.

53 Lahm, H.W., Langen, H. (2000) Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis* **21**, 2105-2114.

54 Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567.

55 Eng, J.K., McCormack, A.L., Yates, J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976 - 989.

56 Shevchenko, A., Wilm, M., Mann, M. (1997) Peptide sequencing by mass spectrometry for homology searches and cloning of genes. *J. Protein Chem.* **16**, 481-490.

57 Mann, M., Wilm, M. (1994) Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* **66**, 4390-4399.

58 Sunyaev, S., Liska, A.J., Golod, A., Shevchenko, A., Shevchenko, A. (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* **75**, 1307-1315.

59 Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W. *et al.* (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI- Quadrupole Time-of-Flight mass spectrometry and BLAST homology searching. *Anal Chem* **73**, 1917-1926.

60 Taylor, J.A., Johnson, R.S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* **73**, 2594-2604.

61 Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. *et al.* (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386.

62 Oda, Y., Huang, K., Cross, F.R., Cowburn, D., Chait, B.T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96**, 6591-6596.

63 Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-999.

64 Chelius, D., Bondarenko, P.V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* **1**, 317-323.

65 Desiderio, D.M., Wirth, U., Lovelace, J.L., Fridland, G., Umstot, E.S., Nguyen, T.M. *et al.* (2000) Matrix-assisted laser desorption/ionization mass spectrometric quantification of the mu opioid receptor agonist DAMGO in ovine plasma. *J Mass Spectrom* **35**, 725-733.

66 Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., Gygi, S.P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100**, 6940-6945.

67 Conrads, T.P., Issaq, H.J., Veenstra, T.D. (2002) New tools for quantitative phosphoproteome analysis. *Biochem Biophys Res Commun* **290**, 885-890.

68 Zhou, H., Ranish, J.A., Watts, J.D., Aebersold, R. (2002) Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nat Biotechnol* **20**, 512-515.

69 Munchbach, M., Quadroni, M., Miotto, G., James, P. (2000) Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal Chem* **72**, 4047-4057.

70 Liu, Y., Patricelli, M.P., Cravatt, B.F. (1999) Activity-based protein profiling: the serine hydrolases. *Proc Natl Acad Sci U S A* **96**, 14694-14699.

71 Greenbaum, D., Medzihradszky, K.F., Burlingame, A., Bogyo, M. (2000) Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem Biol* **7**, 569-581.

72 Zhou, H., Watts, J.D., Aebersold, R. (2001) A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol* **19**, 375-378.

73 Zhang, H., Li, X.J., Martin, D.B., Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* **21**, 660-666.

74 Desiderio, D.M., Kai, M. (1983) Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biomed Mass Spectrom* **10**, 471-479.

75 Mirgorodskaya, O.A., Kozmin, Y.P., Titov, M.I., Korner, R., Sonksen, C.P., Roepstorff, P. (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ ionization mass spectrometry using 18O-labeled internal standards. *Rapid Commun. Mass Spectrom.* **14**, 1226-1232.

76 Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291-294.

77 Fields, S., Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.

78 Lamond, A.I., Mann, M. (1997) Cell biology and the genome projects - A concerted strategy for characterizing multiprotein complexes by using mass spectrometry. *Trends Cell Biol* **7**, 139-142.

79 von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.

80 Hazbun, T.R., Malmstrom, L., Anderson, S., Graczyk, B.J., Fox, B., Riffle, M. *et al.* (2003) Assigning function to yeast proteins by integration of technologies. *Mol Cell* **12**, 1353-1365.

81 Rappsilber, J., Siniossoglou, S., Hurt, E.C., Mann, M. (2000) A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal Chem* **72**, 267-275.

82 Mann, M., Hendrickson, R.C., Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**, 437-473.

83 Smith, D.B., Johnson, K.S. (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* **67**, 31-40.

84 Hoffmann, A., Roeder, R.G. (1991) Purification of his-tagged proteins in non-denaturing conditions suggests a convenient method for protein interaction studies. *Nucleic Acids Res* **19**, 6337-6338.

85 Janknecht, R., de Martynoff, G., Lou, J., Hipskind, R.A., Nordheim, A., Stunnenberg, H.G. (1991) Rapid and efficient purification of native histidine-

tagged protein expressed by recombinant vaccinia virus. *Proc Natl Acad Sci U S A* **88**, 8972-8976.

86 Hopp, T.P., Prickett, K.S., Price, V.L., Libby, R.T., March, C.J., Cerretti, D.P. *et al.* (1988) A short polypeptide marker sequence useful for recombinant protein identification and purification. *Bio-Technol* **6**, 1204-1210.

87 Phillips, T.M., Queen, W.D., More, N.S., Thompson, A.M. (1985) Protein A-coated glass beads. Universal support medium for high-performance immunoaffinity chromatography. *J Chromatogr* **327**, 213-219.

88 Graumann, J., Dunipace, L.A., Seol, J.H., McDonald, W.H., Yates, J.R., 3rd, Wold, B.J. *et al.* (2003) Applicability of TAP-MudPIT to pathway proteomics in yeast. *Mol Cell Proteomics*.

89 Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**, 1030-1032.

90 Phizicky, E., Bastiaens, P.I., Zhu, H., Snyder, M., Fields, S. (2003) Protein analysis on a proteomic scale. *Nature* **422**, 208-215.

91 Ito, T., Ota, K., Kubota, H., Yamaguchi, Y., Chiba, T., Sakuraba, K. *et al.* (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol Cell Proteomics* **1**, 561-566.

92 Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J.C., Colley, A., Dix, I. *et al.* (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17**, 95-110.

93 Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisia*e. *Nature* **403**, 623-627.

94 Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-4574.

95 Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S. *et al.* (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211-215.

96 Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116-122.

97 Wouters, F.S., Verveer, P.J., Bastiaens, P.I. (2001) Imaging biochemistry inside cells. *Trends Cell Biol* **11**, 203-211.

98 Siegel, R.M., Frederiksen, J.K., Zacharias, D.A., Chan, F.K., Johnson, M., Lynch, D. *et al.* (2000) Fas preassociation required for apoptosis signaling and dominant inhibition by pathogenic mutations. *Science* **288**, 2354-2357.

99 Mahajan, N.P., Linder, K., Berry, G., Gordon, G.W., Heim, R., Herman, B. (1998) Bcl-2 and Bax interactions in mitochondria probed with green fluorescent protein and fluorescence resonance energy transfer. *Nat Biotechnol* **16**, 547-552.

100 Day, R.N. (1998) Visualization of Pit-1 transcription factor interactions in the living cell nucleus by fluorescence resonance energy transfer microscopy. *Mol Endocrinol* **12**, 1410-1419.

101 Sorkin, A., McClure, M., Huang, F., Carter, R. (2000) Interaction of EGF receptor and grb2 in living cells visualized by fluorescence resonance energy transfer (FRET) microscopy. *Curr Biol* **10**, 1395-1398.

102 Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-2368.

103 Huynen, M., Snel, B., Lathe, W., 3rd, Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10**, 1204-1210.

104 Dandekar, T., Snel, B., Huynen, M., Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-328.

105 Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288.

106 Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

107 Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303-305.

108 Bader, G.D., Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2.

109 Duan, X.J., Xenarios, I., Eisenberg, D. (2002) Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. *Mol Cell Proteomics* **1**, 104-116.

110 Csank, C., Costanzo, M.C., Hirschman, J., Hodges, P., Kranz, J.E., Mangan, M. *et al.* (2002) Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol* **350**, 347-373.

111 Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-34.

112 Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G. *et al.* (2002) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res* **30**, 91-93.

113 Gharahdaghi, F., Weinberg, C.R., Meagher, D.A., Imai, B.S., Mische, S.M. (1999) Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: a method for the removal of silver ions to enhance sensitivity. *Electrophoresis* **20**, 601-605.

114 Fernandez-Patron, C., Calero, M., Collazo, P.R., Garcia, J.R., Madrazo, J., Musacchio, A. *et al.* (1995) Protein reverse staining: high-efficiency microanalysis of unmodified proteins detected on electrophoresis gels. *Anal Biochem* **224**, 203-211.

115 Havlis, J., Thomas, H., Sebela, M., Shevchenko, A. (2003) Fast responce proteomics by accelerated in-gel digestion of proteins. *Anal Chem* **75**, 1300-1306.

116 Harris, W.A., Janecki, D.J., Reilly, J. (2002) Use of matrix clusters and trypsin autolysis fragments as mass calibrants in matrix-assisted laser desorption/

ionization time-of-flight mass spectrometer. *Rapid Comm Mass Spectrom* **16**, 1714-1722.

117 Shevchenko, A., Chernushevic, I., Wilm, M., Mann, M. (2002) "*De novo*" sequencing of peptides recovered from in-gel digested proteins by nanoelectrospray tandem mass spectrometry. *Mol Biotechnol* **20**, 107-118.

118 Liska, A.J., Shevchenko, A. (2003) Combining mass spectrometry with database interrogation strategies in proteomics. *Trends Anal Chem* **22**, 291-298.

119 Shevchenko, A., Sunyaev, S., Liska, A., Bork, P., Shevchenko, A. (2002) Nanoelectrospray tandem mass spectrometry and sequence similarity searching for identification of proteins from organisms with unknown genomes. *Meth Mol Biol* **211**, 221-234.

120 Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E. *et al.* (2001) The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods* **24**, 218-229.

121 Krawchuk, M.D., Wahls, W.P. (1999) High-efficiency gene targeting in Schizosaccharomyces pombe using a modular, PCR-based approach with long tracts of flanking homology. *Yeast* **15**, 1419-1427.

122 Roguev, A., Schaft, D., Shevchenko, A., Aasland, R., Stewart, A.F. (2003) High conservation of the Set1/Rad6 axis of histone 3 lysine 4 methylation in budding and fission yeasts. *J Biol Chem* **278**, 8487-8493.

123 Roguev, A., Shevchenko, A., Schaft, D., Thomas, H., Stewart, A.F. (2004) A comparative analysis of an orthologous proteomic environment in the yeasts *S.cerevisiae* and *S.pombe*. *Mol Cell Proteomics* **3**, *in press*.

124 Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O. *et al.* (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 14440-14445.

125 Krause, E., Wenschuh, H., Jungblut, P.R. (1999) The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal Chem* **71**, 4160-4165.

126 Tabb, D.L., Smith, L.L., Breci, L.A., Wysocki, V.H., Lin, D., Yates, J.R., 3rd (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* **75**, 1155-1163.

127 Schnölzer, M., Jedrzejewski, P., Lehmann, W.D. (1996) Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. *Electrophoresis* **17**, 945-953.

128 Shevchenko, A., Wilm, M., Shevchenko, A. (2000) Quantitative evaluation of the efficiency of in-gel digestion of proteins by isotopic labeling and mass spectrometry. *Proc. 48th ASMS Conf. Mass Spectrom. Allied Topics, Long Beach CA*, 859-860.

129 Shevchenko, A., Shevchenko, A. (2001) Evaluation of the efficiency of in-gel digestion of proteins by peptide isotopic labeling and maldi mass spectrometry. *Anal Biochem* **296**, 279-283.

130 Yao, X., Freas, A., Ramirez, J., Demirev, P.A., Fenselau, C. (2001) Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842.

131 Reynolds, K., Yao, X., Fenselau, C. (2002) Proteolytic O-18 labeling for comparative proteomics: Evaluation of endoprotease Glu-C as the catalytic agent. *J Proteome Res* **1**, 27-33.

132 Parker, K.C., Garrels, J.I., Hines, W., Butler, E.M., McKee, A.H., Patterson, D. *et al.* (1998) Identification of yeast proteins from two-dimensional gels: working out spot cross-contamination. *Electrophoresis* **19**, 1920-1932.

133 Borchers, C., Peter, J.F., Hall, M.C., Kunkel, T.A., Tomer, K.B. (2000) Identification of *in-gel* digested proteins by complementary peptide mass fingerprinting and tandem mass spectrometry data obtained on an electrospray ionization quadrupole time-of-flight mass spectrometer. *Anal. Chem.* **72**, 1163-1168.

134 Shevchenko, A., Loboda, A., Ens, W., Schraven, B., Standing, K.G., Shevchenko, A. (2001) Archived polyarylamide gels as a resource for proteome characterization by mass spectrometry. *Electrophoresis* **22**, 1194-1203.

135 Medzihradszky, K.F., Leffler, H., Baldwin, M.A., Burlingame, A.L. (2001) Protein identification by in-gel digestion, high-performance liquid chromatography, and mass spectrometry: peptide analysis by complementary ionization techniques. *J Am Soc Mass Spectrom* **12**, 215-221.

136 Shevchenko, A., Schaft, D., Roguev, A., Pijnappel, W.W.M.P., Stewart, A.F., Shevchenko, A. (2002) Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry: analytical perspective. *Mol Cell Proteomics* **1**, 204-212.

137 Roguev, A., Schaft, D., Shevchenko, A., Pijnappel, W.W.M., Wilm, M., Aasland, R. *et al.* (2001) The *S. cerevisiae* Set1 complex includes an Ash2 homolog and methylates histone 3 lysine 4. *EMBO J* **20**, 7137-7148.

138 Shevchenko, A., Shevchenko, A., Schraven, B., Mann, M. (1998) Looking back into the future: microsequencing of proteins from archived gels stored for years. *Proc. 46th ASMS Conf. Mass Spectrom. and Allied Topics, Orlando FL*, 238.

139 Matsumoto, H., Komori, N. (1999) Protein identification on two-dimensional gels archived nearly two decades ago by in-gel digestion and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Anal Biochem* **270**, 176-179.

140 Schraven, B., Schoenhaut, D., Bruyns, E., Koretzky, G., Eckerskorn, C., Wallich, R. *et al.* (1994) LPAP, a novel 32-kDa phosphoprotein that interacts with CD45 in human lymphocytes. *J Biol Chem* **269**, 29102-29111.

141 Ainsztein, A.M., Kandels-Lewis, S.E., Mackay, A.M., Earnshaw, W.C. (1998) INCENP centromere and spindle targeting: identification of essential conserved motifs and involvement of heterochromatin protein HP1. *J Cell Biol* **143**, 1763-1774.

142 Consortium, I.H.G.S. (2001) *Nature* **409**, 860-921.

143 Venter, J.C. (2001) The sequence of the human genome. *Science* **291**, 1304-1351.

144 Liska, A.J., Shevchenko, A. (2003) Expanding organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics* **3**, 19-28.

145 Clauser, K.R., Hall, S.C., Smith, D.M., Webb, J.W., Andrews, L.E., Tran, H.M. *et al.* (1995) Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. *Proc Natl Acad Sci U S A* **92**, 5072-5076.

146 Swiderek, K.M., Davis, M.T., Lee, T.D. (1998) The identification of peptide modifications derived from gel-separated proteins using electrospray triple quadrupole and ion trap analyses. *Electrophoresis* **19**, 989-997.

147 Thiede, B., Lamer, S., Mattow, J., Siejak, F., Dimmler, C., Rudel, T. *et al.* (2000) Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid. Commun. Mass Spectrom.* **14**, 496-502.

148 Sheibani, N. (1999) Prokaryotic gene fusion expression systems and their use in structural and functional studies of proteins. *Prep Biochem Biotechnol* **29**, 77-90.

149 Ford, C.F., Suominen, I., Glatz, C.E. (1991) Fusion tails for the recovery and purification of recombinant proteins. *Protein Expr Purif* **2**, 95-107.

150 Gottschalk, A., Tang, J., Puig, O., Salgado, J., Neubauer, G., Colot, H.V. *et al.* (1998) A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA* **4**, 374-393.

151 Riedel, N., Wise, J.A., Swerdlow, H., Mak, A., Guthrie, C. (1986) Small nuclear RNAs from *Saccharomyces cerevisiae*: unexpected diversity in abundance, size, and molecular complexity. *Proc Natl Acad Sci U S A* **83**, 8097-8101.

152 Smith, E.R., Eisen, A., Gu, W., Sattah, M., Pannuti, A., Zhou, J. *et al.* (1998) ESA1 is a histone acetyltransferase that is essential for growth in yeast. *Proc Natl Acad Sci U S A* **95**, 3561-3565.

153 Galarneau, L., Nourani, A., Boudreault, A.A., Zhang, Y., Heliot, L., Allard, S. *et al.* (2000) Multiple links between the NuA4 histone acetyltransferase complex and epigenetic control of transcription. *Mol Cell* **5**, 927-937.

154 Schaft, D., Roguev, A., Kotovic, K.M., Shevchenko, A., Sarov, M., Shevchenko, A. *et al.* (2003) The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation. *Nucleic Acid Res* **31**, 2475-2482.

155 Pijnappel, W.W., Schaft, D., Roguev, A., Shevchenko, A., Tekotte, H., Wilm, M. *et al.* (2001) The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes Dev* **15**, 2991-3004.

156 Caspary, F., Shevchenko, A., Wilm, M., Seraphin, B. (1999) Partial purification of the yeast U2 snRNP reveals a novel yeast pre- mRNA splicing factor required for pre-spliceosome assembly. *Embo J* **18**, 3463-3474.

157 Bouveret, E., Rigaut, G., Shevchenko, A., Wilm, M., Seraphin, B. (2000) A Sm-like protein complex that participates in mRNA degradation. *EMBO J* **19**, 1661-1671.

158 Shou, W., Seol, J.H., Shevchenko, A., Baskerville, C., Moazed, D., Chen, Z.W. *et al.* (1999) Exit from mitosis is triggered by Tem1-dependent release of the protein phosphatase Cdc14 from nucleolar RENT complex. *Cell* **97**, 233-244.

159 Shevchenko, A., Mann, M. in Mass spectrometry in biology and medicine. (eds. A.L. Burlingame, S.A. Carr & M. Baldwin) 237-269 (Humana Press, Totowa, NJ; 1999).

160 Kurland, C.G. (1991) Codon bias and gene expression. *FEBS Lett* **285**, 165-169.

161 Sharp, P.M., Li, W.H. (1987) The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-1295.

162 Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **97**, 9390-9395.

163 Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S. *et al.* (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* **29**, 75-79.

164 Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H. *et al.* (1996) Live with 6000 Genes. *Science* **274**, 546 - 567.

165 Goffeau, A. (2000) Four years of post-genomic life with 6000 yeast genes. *FEBS Lett* **480**, 37-41.

166 Hodges, P.E., McKee, A.H.Z., Davis, B.P., Payne, W.E., Garrels, J.L. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res* **27**, 69-73.

167 Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-753.

168 Krogan, N.J., Kim, M., Ahn, S.H., Zhong, G., Kobor, M.S., Cagney, G. *et al.* (2002) RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted proteomics approach. *Mol Cell Biol* **22**, 6979-6992.

169 Evangelista, C., Lockshon, D., Fields, S. (1996) The yeast two-hybrid system: prospects for protein linkage maps. *Trends Cell Biol* **6**, 196 -199.

170 Seol, J.H., Shevchenko, A., Shevchenko, A., Deshaies, R.J. (2001) Skp1 forms multiple protein complexes, including RAVE, a regulator of V- ATPase assembly. *Nat Cell Biol* **3**, 384-391.

171 Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A. *et al.* (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133-1143.

172 Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. *et al.* (2002) Comparative assessment of large-scale data sets of protein protein interactions. *Nature* **417**, 399-403.

173 Shevchenko, A., Zachariae, W., Shevchenko, A. (1999) A strategy for the characterization of protein interaction networks by mass spectrometry. *Biochem. Soc. Trans.* **27**, 549-554.

174 Deshaies, R.J., Seol, J.H., McDonald, W.H., Cope, G., Lyapina, S., Shevchenko, A. *et al.* (2002) Charting the protein complexome in yeast by mass spectrometry. *Mol Cell Proteomics* **1**, 3-10.

175 Ivanova, A.V., Bonaduce, M.J., Ivanov, S.V., Klar, A.J. (1998) The chromo and SET domains of the Clr4 protein are essential for silencing in fission yeast. *Nat Genet* **19**, 192-195.

176 Nislow, C., Ray, E., Pillus, L. (1997) SET1, a yeast member of the trithorax family, functions in transcriptional silencing and diverse cellular processes. *Mol Biol Cell* **8**, 2421-2436.

177 Capili, A.D., Schultz, D.C., Rauscher, I.F., Borden, K.L. (2001) Solution structure of the PHD domain from the KAP-1 corepressor: structural determinants for PHD, RING and LIM zinc-binding domains. *Embo J* **20**, 165-177.

178 Gilchrist, D., Mykytka, B., Rexach, M. (2002) Accelerating the rate of disassembly of karyopherin.cargo complexes. *J Biol Chem* **277**, 18161-18172.

179 Rundlett, S.E., Carmen, A.A., Kobayashi, R., Bavykin, S., Turner, B.M., Grunstein, M. (1996) HDA1 and RPD3 are members of distinct yeast histone deacetylase complexes that regulate silencing and transcription. *Proc Natl Acad Sci U S A* **93**, 14503-14508.

180 Brachmann, C.B., Sherman, J.M., Devine, S.E., Cameron, E.E., Pillus, L., Boeke, J.D. (1995) The SIR2 gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression, and chromosome stability. *Genes Dev* **9**, 2888-2902.

181 Cockell, M., Renauld, H., Watt, P., Gasser, S.M. (1998) Sif2p interacts with Sir4p amino-terminal domain and antagonizes telomeric silencing in yeast. *Curr Biol* **8**, 787-790.

182 Aasland, R., Stewart, A.F., Gibson, T. (1996) The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB. *Trends Biochem Sci* **21**, 87-88.

183 Dolinski, K., Heitman, J. in Guidebook to molecular chaperones and protein folding catalysts. (ed. S. Tooze) 359-369 (Oxford University Press, Oxford, UK; 1997).

184 Stoldt, V., Rademacher, F., Kehren, V., Ernst, J.F., Pearce, D.A., Sherman, F. (1996) Review: the Cct eukaryotic chaperonin subunits of *Saccharomyces cerevisiae* and other yeasts. *Yeast* **12**, 523-529.

185 Krogan, N.J., Dover, J., Khorrami, S., Greenblatt, J.F., Schneider, J., Johnston, M. *et al.* (2002) COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* **277**, 10753-10755.

186 Dichtl, B., Blank, D., Ohnacker, M., Friedlein, A., Roeder, D., Langen, H. *et al.* (2002) A role for SSU72 in balancing RNA polymerase II transcription elongation and termination. *Mol Cell* **10**, 1139-1150.

187 Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B Biol Sci* **256**, 119-124.

188 Lynch, M., Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155.

189 Wysocka, J., Myers, M.P., Laherty, C.D., Eisenman, R.N., Herr, W. (2003) Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell- proliferation factor HCF-1. *Genes Dev* **17**, 896-911.

Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel anfertigt habe. Die aus fremden Quellen direkt oder indirekt ubernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bishier weder im Inland noch im Ausland in gleicher oder änlicher form einer anderen Prüfungsbehörde vorgelegt.

Die Promotionsordnung wird anerkannt.

Anna Shevchenko.

AKNOWLEDGEMENTS

I would also like to thank all my collaborators and co-authors for insight, suggestions, criticisms and support. Most of all, I am thankful to the exceptional staff of the laboratories of Prof. Dr. A. F. Stewart (EMBL / TU Dresden), Dr. B. Seraphin (EMBL), who helped me to understand the field of proteomics and interactomics from the molecular biology perspective; laboratories Dr. Matthias Mann (EMBL) and Dr. Matthias Wilm (EMBL), where I learned mass spectrometry and discovered how spectacular the proteomics research can be; laboratory of Dr. Andrej Shevchenko (MPI CBG) for continuous support of my research endeavors.

I am grateful to Igor Chernushevich and Alexander Loboda (MDS Sciex, Canada) as well as to members of Prof. Dr. Kenneth Standing laboratory in the University of Manitoba for sharing their expertise on quadrupole time-of-flight mass spectrometry.

I am also grateful to Prof. Dr. Gerhard Roedel (TU Dresden) for his expert advice and guidance through the thesis work.